Polarity Analysis of Micro Reviews in Foursquare

Felipe Moraes, Marisa Vasconcelos, Patrick Prado
Jussara Almeida e Marcos Gonçalves
{felipemoraes,marisav,patrickprado,jussara,mgoncalv}@dcc.ufmg.br
Universidade Federal de Minas Gerais, Belo Horizonte, Brasil

ABSTRACT

On Foursquare, one of the currently most popular location-based social networks, users can not only share which places (venues) they visit but also leave short comments (tips) about their previous experiences at specific venues. Tips may provide a valuable feedback for business owners as well as for potential new customers. Sentiment or polarity classification provides useful tools for opinion summarization, which can help both parties to quickly obtain a predominant view of the opinions posted by users at a specific venue. We here present what, to our knowledge, is the first study of polarity of Foursquare tips. We start by characterizing two datasets of collected tips with respect to their textual content. Some inherent characteristics of tips, such as short sizes as well as informal and often noisy content, pose great challenges to polarity detection. We then investigate the effectiveness of four alternative polarity classification strategies on subsets of our dataset. Three of the considered strategies are based on supervised machine learning techniques and the fourth one is an unsupervised lexicon-based approach. Our evaluation indicates that effective polarity classification can be achieved even if the simpler lexicon-based approach, which does not require costly manual tip labeling, is adopted.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Web 2.0 applications, Sentiment Analysis, Micro-reviews

1. INTRODUÇÃO

A popularização do acesso aos *smartphones*, a disponibilidade de tecnologias de geolocalização como GPS (*Global Positioning System*) e o crescente interesse nas redes sociais possibilitaram o surgimento de redes sociais baseadas em geolocalização (LBSN do inglês *Location-Based Social Network*), como o Foursquare e redes que se utilizam de serviços de geolocalização como o Google Plus e o Instagram[1] dentre outras. Com mais de 30 milhões de usuários, o Foursquare é a LBSN mais popular atualmente. Nessa rede social, os usuários podem, além de compartilhar a sua localização em uma variedade de locais (*venues*), compartilhar suas experiências e opiniões através de micro-revisões ou *tips*.

As *tips* são textos curtos e informais adicionados por usuários do Foursquare a respeito de algum local (*venue*) já frequentado por eles. Geralmente, elas provêem algum tipo de recomendação sobre o local ou serviço oferecido, mas podem também conter críticas ou reclamações. Além disso, os usuários também podem curtir (*like*) uma *tip*, demonstrando assim seu interesse e fazendo com que essa *tip* fique mais visível para o resto da comunidade. Além de úteis para os próprios usuários da rede, as *tips* também são essenciais para os proprietários desses locais, já que esse *feedback* pode ajudá-los a aperfeiçoar o funcionamento do próprio negócio.

Este trabalho analisa métodos para detecção da polaridade ou sentimento em micro-revisões em duas bases de *tips* coletadas do Foursquare. A análise de polaridade tem o objetivo de identificar automaticamente a atitude do autor a respeito de um determinado tópico ou o tom utilizado no texto inteiro, classificando esse conteúdo em positivo ou negativo. Métodos propostos para detecção de polaridade possuem diversas aplicações que vão desde sumarização de opiniões, por exemplo, em revisões online [19] até aplicações que monitoram a opinião das pessoas em tempo real [6]. No contexto do Foursquare, a identificação ou detecção da polaridade de uma *tip* pode ser usada para sumarizar o sentimento de diversos usuários sobre um local e fornecer ao proprietário uma visão geral sobre o que estão falando sobre o seu empreendimento. Além disso, os próprios usuários podem se beneficiar desse tipo de informação para os auxiliar na escolha de um local para visitar. Contudo, o Foursquare traz desafios próprios no que concerne a detecção de polaridade das micro-revisões. Em particular, *tips* são geralmente curtas (limitadas a 200 caracteres) e bastante informais, o que significa que a informação coletada pode não ser suficiente para detecção da polaridade do texto. Além disso, gírias e expressões (por exemplo, "legalll!!!") comumente encontradas em redes sociais, dificultam muito a análise.

As técnicas de detecção de polaridade existentes podem ser agrupadas em *supervisionadas* e *não supervisionadas*. Nas técnicas supervisionadas, classificadores automáticos são treinados a partir de exemplos que podem ser manualmente rotulados ou obtidos de alguma outra base rotulada disponível [1, 3, 5, 16, 17, 19]. Dentre as técnicas não supervisionadas, as mais utilizadas são aquelas baseadas em listas de palavras positivas e negativas (léxico), em que a classificação é feita considerando a polaridade de cada termo do texto [3, 5, 16, 17]. Cada uma dessas abordagens possui vanta-

WebMedia'13, November 5–8, 2013, Salvador, Brazil.

http://dx.doi.org/10.1145/2526188.2526195 .

[1] `http://foursquare.com/`, `http://plus.google.com/` e `https://instagram.com/`

gens e desvantagens. A abordagem supervisionada apesar de eficaz para classificação de polaridade em ambientes "tradicionais" [18], é muito dependente do conjunto de dados rotulados para o treinamento dos modelos. Como esses dados são rotulados por pessoas, esses rótulos também podem estar sujeitos a erros de classificação devido à interpretação do texto. Além disso, o número de instâncias de treino afeta o desempenho da classificação em termos da acurácia, que está diretamente relacionada à cobertura do maior número de cenários possíveis [12], o que implica em custos altos para geração de grandes bases de treinamento. Esses problemas são minimizados com a utilização de abordagens que fazem uso de léxicos que não requerem nenhum treinamento para classificação da polaridade dos dados e podem ser aplicados em uma variedade de contextos ou aplicações. Contudo, não há léxicos disponíveis que sejam totalmente adequados para todos os domínios de aplicação (e.g., gírias comuns nas redes sociais estão ausentes na maioria dos léxicos [9]). Isso se deve a fato que os usuários podem utilizar palavras cujos sentimentos são sensíveis ao tópico do domínio [12]. Mais ainda, para textos curtos como as micro-revisões, não há um consenso sobre a melhor abordagem [5, 3, 17]. Essa falta de consenso é um dos principais motivadores de nosso estudo.

Neste contexto, este artigo avalia a eficácia de quatro técnicas diferentes para detecção automática da polaridade de *tips* no Foursquare. São consideradas tanto técnicas supervisionadas quanto não supervisionadas. As abordagens supervisionadas analisadas são três classificadores comumente utilizados no contexto de análise de sentimento: Naïve Bayes, Máxima Entropia e SVM (*Support Vector Machine*) [11, 15, 24]. Como abordagem não supervisionada foi proposta e analisada uma técnica baseada no léxico SentiWordNet [4].

Para avaliação dessas abordagens, foram utilizadas duas bases de *tips* coletadas do Foursquare: uma base rotulada manualmente por voluntários e outra base com *tips* contendo pelo menos um emoticon[2]. As duas bases foram caracterizadas em relação ao conteúdo textual de suas *tips* e a localização geográfica do local onde as *tips* foram postadas. Os resultados da avaliação apontam que a abordagem não supervisionada (léxico) produz resultados comparáveis aos dos métodos supervisionados, sem os custos de rotulação. Foi observado também que em ambas as bases, a abordagem não supervisionada (léxico) é a melhor para se detectar *tips positivas*, enquanto que para detecção de *tips negativas*, o melhor método é o classificador Naïve Bayes, que, dentre os métodos supervisionados, foi o que apresentou melhor eficácia.

O restante deste artigo está organizado como segue. A Seção 2 discute trabalhos relacionados, e a Seção 3 descreve as bases de dados usadas. A Seção 4 apresenta uma breve análise das principais características das *tips* e sua distribuição geográfica. As técnicas de detecção de polaridade analisadas são apresentadas na Seção 5, enquanto os resultados desta avaliação são discutidos na Seção 6. Finalmente, a Seção 7 apresenta as conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

Na literatura existem vários trabalhos sobre análise de sentimento, principalmente de textos longos, mais especificamente revisões longas [16, 19]. Porém, trabalhos que analisam textos curtos como as *tips* do Foursquare ainda compõem uma área recente de investigação. Todos esses trabalhos podem ser agrupados em duas principais abordagens: métodos supervisionados [1, 3, 5, 17] e métodos não supervisionados baseados em léxicos [3, 5, 17].

[2]Emoticons são sequências de caracteres tais como :) que representam expressões faciais.

No contexto de revisões online, os principais trabalhos de análise de polaridade ou sentimento focam em revisões longas. Por exemplo, Pang *et al.* [19] comparam vários algoritmos de classificação supervisionados (Naïve Bayes, Máxima Entropia e SVM) para detecção de sentimento em revisões de filmes. Cada revisão foi representada como um *bag-of-words* baseada em unigramas e bigramas, e a melhor acurácia foi obtida pelo método SVM utilizando unigramas. Ohana *et al.* utilizam o léxico SentiWordNet também para a classificação automática da polaridade em revisões de filme. Os autores propõem duas abordagens para a realização dessa tarefa. Na primeira abordagem, o léxico foi utilizado para contabilizar a pontuação das palavras positivas e negativas do texto. A polaridade da revisão era determinada pela polaridade com maior pontuação. Na segunda abordagem, os autores utilizaram as pontuações positivas e negativas para o treinamento de um modelo supervisionado baseado no SVM. Para o presente estudo, cujo foco é em textos curtos, nós utilizamos um método não supervisionado semelhante à primeira abordagem proposta por Ohana *et al.* [16] para detecção da polaridade de *tips* no Foursquare. A representação dos termos de uma *tip* também foi modelada como uma *bag-of-words*. No entanto, nossos resultados foram melhores com a utilização de TF-IDF (i.e., produto do *Term Frequency* pelo *Inverse Document Frequency* [2]), ao invés do uso de unigramas ou bigramas.

Os trabalhos que analisam a polaridade de textos curtos focam majoritariamente em redes de micro-blogs como o Twitter [1, 5, 17]. Em Aisopos *et al.* [1], os autores propõem um método para classificação de *tweets* em positivos e negativos baseado em evidências textuais e evidências relacionadas ao contexto social do autor do *tweet* (tamanho da rede social e polaridade das mensagens dos seguidores). Assim, os atributos foram modelados pelo TF-IDF e o rótulo de um *tweet* foi determinado pela presença de emoticons positivos ou negativos, tal qual a abordagem adotada aqui para rotulagem de uma das bases analisadas.

Alguns poucos trabalhos realizaram comparações entre as abordagens supervisionadas e não supervisionada para classificação da polaridade em textos curtos [5, 3, 17]. Em Go *et al.* [5], os autores demonstram que os mesmos métodos supervisionados propostos por Pang *et al.* [19] para revisões longas podem ser aplicados a textos curtos como *tweets* ao comparar a acurácia desses métodos com um método não supervisionado baseado em um léxico extraído do site *Twittratr*[3]. Em Bermingham *et al.* [3], os autores concluíram que os métodos supervisionados (SVM e Multinominal Naïve Bayes) apresentaram melhor desempenho, em termos de acurácia, que um método não supervisionado baseado no SentiWordnet no contexto de *tweets* e microrevisões do site Blippr. Já em Paltoglou *et al.* [17] os autores propõem um método não supervisionado baseado no léxico LIWC [21] para classificação de sentimento de *tweets* e comentários no MySpace e no Digg, e mostram que o método proposto supera, em termos de acurácia, três métodos supervisionados (Naïve Bayes, Máxima Entropia e SVM).

Portanto, pode-se observar que não há um consenso entre os trabalhos [5, 3, 17] sobre o melhor tipo de abordagem a ser aplicada em textos curtos. Nosso trabalho pretende contribuir com essa discussão, focando na análise de polaridade de *tips* no Foursquare, um ambiente "ruidoso" mas rico de informação. Até onde sabemos, este é o primeiro trabalho que investiga polaridade de sentimento neste contexto.

3. BASES DE DADOS

O Foursquare é a maior e a mais popular rede social baseada em geolocalização (LBSN), atualmente com mais de 30 milhões

[3]http://twitrratr.com/.

de usuários pelo mundo[4]. As aplicações desenvolvidas para serviços de LBSN são baseadas na tecnologia GPS (Global Positioning System) que permite aos usuários interagir, compartilhar e recomendar locais (*venues*) baseados em sua localização atual. Os *venues* são representações virtuais de locais do mundo real, como escolas, lojas, restaurantes ou aeroportos, entre outros. O Foursquare classifica os *venues* em 9 categorias pré-definidas: "Food", "Travel & Transport", "Great Outdoors", "Nightlife Spots","Professional& Other Places", "Residences", "Shops& Services", "Colleges & Universities"e "Arts & Entertainment". Para indicar o local onde o usuário se encontra ele deve realizar um *check in*, informando seus amigos da rede social onde ele se encontra naquele momento.

Além de compartilhar locais, os usuários podem realizar recomendações, adicionando comentários sobre esses locais através de *tips*. *Tips* são como resenhas, porém mais concisas e limitadas a 200 caracteres, que podem ter um caráter informativo, recomendativo ou descritivo, relatando experiências dos usuários. As *tips* podem ser avaliadas por outros usuários, que podem curtí-las (*like*) tornando-as mais visíveis à comunidade do Foursquare.

A avaliação das abordagens de detecção de polaridade de uma *tip* foi feita utilizando duas bases de *tips* coletadas do Foursquare: uma rotulada manualmente por voluntários e outra composta por *tips* com pelo menos 1 emoticon. As duas bases são subconjuntos de uma base de aproximadamente 10 milhões de *tips* postadas por 13 milhões de usuários coletada de Agosto a Outubro de 2011 usando a API do Foursquare. Além disso, na construção das duas bases (manual e emoticon) foram consideradas somente *tips* postadas em *venues* onde a língua oficial[5] é a língua inglesa[6].

A **base manualmente rotulada** (ou simplesmente base manual) foi construída da seguinte maneira. Foram selecionadas aleatoriamente 1.250 *tips* para serem rotuladas por 15 voluntários. Cada *tip* foi analisada por um grupo de três voluntários evitando assim a ocorrência de empates. Para cada avaliador voluntário, foi apresentado uma amostra de 250 *tips* e algumas informações, como o nome e a categoria do *venue* onde foi postada a *tip*. Foi pedido aos voluntários que rotulassem o conteúdo de cada uma das *tips* como positiva, negativa ou neutra. Foi observado que os voluntários concordaram em 94% das *tips* selecionadas[7], resultando na seguinte classificação: 57,78% das *tips* foram classificadas como positivas, 15,64% como negativa enquanto que 26,58% foram consideradas como sendo neutras. Como em [1, 19], este trabalho foca na classificação entre as polaridades positiva e negativa. Desta forma, as *tips* neutras foram desconsideradas para nossas análises. Finalmente, para essa base foram consideradas apenas *tips* com pelo menos 1 palavra no SentiWordNet. Após esta filtragem, foram mantidas 851 *tips* para constituir a base manualmente rotulada.

Já para a **base com emoticons**, foram selecionadas 3.512 *tips* com pelo menos 1 emoticon dentre todas as *tips* em inglês. Emoticons podem servir com rótulos ruidosos (*noisy labels*), em que o símbolo :) em uma *tip* pode indicar que a *tip* contenha uma polaridade positiva, enquanto que :(pode indicar uma polaridade negativa [20]. São chamados de ruidosos porque certos textos não são facilmente classificáveis como, por exemplo, aqueles que expressam sarcasmo. Como em [1, 5], foi assumido que emoticons positivos (:)', '(:', ':-)', '(-:', ':)', ':D', '=D' indicam *tips* positivas enquanto que emoticons negativos ':(', '):', ':-(', ')-:', ': (', ') :'. indicam *tips* negativas. Para que as tips pudessem ser consideradas pelo método não supervisionado, foram consideradas, em ambas as bases, somente tips com pelo menos uma palavra no SentiWordNet. Com a aplicação desse filtro, foi removida apenas uma pequena fração de tips (até 1,6%) de ambas as bases.

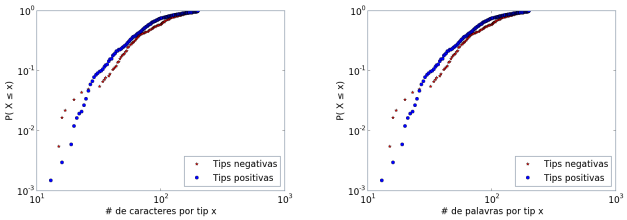
A Tabela 1 sumariza as principais características das duas bases de *tips* utilizadas em nossos experimentos na Seção 6.

Table 1: Sumário das bases de tips utilizadas nos experimentos

	Número de Tips		
	Positivas	**Negativas**	**Total**
Tips manualmente rotuladas	669	182	851
Tips com emoticons	3014	440	3.454

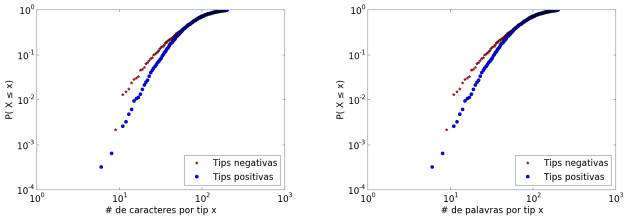
4. ANÁLISE DAS BASES DE TIPS

Nesta seção, são caracterizadas as duas bases de dados utilizadas na avaliação dos métodos de detecção da polaridade de *tips* no Foursquare. Três tipos de atributos foram considerados: textual, tópico discutidos pelas *tips* e a distribuição geográfica dos *venues* aos quais as *tips* foram associadas.



(a) Número de caracteres por tip (b) Número de palavras por tip

Figure 1: Tamanho das tips da base manualmente rotulada



(a) Número de caracteres por tip (b) Número de palavras por tip

Figure 2: Tamanho das tips da base de emoticons

As Figuras 1 e 2 mostram as funções de distribuição de probabilidade acumulada (CDF) do número de caracteres e do número de palavras por *tip* positiva e por *tip negativa* para as bases manual e de emoticons, respectivamente. Ambos os eixos estão em escala logarítmica. Os dois grupos de gráficos mostram pouca diferença entre *tips* positivas e negativas. Entretanto, na base manual, há uma maior tendência para *tips* negativas serem mais longas. De fato, as *tips* negativas nesta base têm em média 93 caracteres e 17 palavras,

[4]http://foursquare.com/about

[5]`http://en.wikipedia.org/wiki/List_of_official_languages`

[6]Essa restrição foi feita já que as ferramentas utilizadas pelo método não supervisionado para computação da polaridade são restritas à língua inglesa.

[7]Pelo menos dois voluntários concordaram com a mesma polaridade. 6% da tips avaliadas foram descartadas, porque não houve consenso entre os avaliadores.

enquanto que as *tips* positivas têm em média 82 caracteres e 15 palavras. Isto sugere que os usuários do Foursquare tendem a ser mais detalhistas ao expressar algum tipo de experiência negativa. Essa característica das *tips* negativas por ser útil para o dono do estabelecimento, que pode ter um *feedback* mais detalhado sobre pontos negativos do seu negócio. Já para a base de emoticons, as diferenças são menores, com tamanhos intermediários (entre 10 e 100 caracteres) sendo ligeiramente mais frequentes entre as *tips* positivas. Em média, as *tips* negativas têm 81 caracteres e 16 palavras enquanto as positivas têm em torno de 83 caracteres e 16 palavras. Esta pequena diferença entre as duas bases pode ser consequência do maior ruído na rotulação da base de emoticon.

Além do tamanho das *tips*, foi também analisada a frequência de diferentes palavras nas *tips* positivas e negativas das duas bases. A Figura 3 mostra as nuvens de palavras para *tips* positivas nas duas bases. A Figura 4 mostra nuvens equivalentes para *tips* negativas. Para as *tips* positivas, podemos notar que os adjetivos *great*, *best* e *good* são os mais frequentes em ambas as bases. Já para as *tips* negativas, os termos mais frequentes são substantivos como *place*, *food* e *service*. Note a presença de palavras positivas como *good* e *best* nas *tips* negativas, que mostram a importância de outros termos negativo como *not* para determinação do sentimento da *tip*. Este aspecto foi tratado na determinação da polaridade da *tip* pela abordagem não supervisionada baseada em léxico. Comparando as duas bases, podemos observar que palavras relacionadas a bares e restaurantes como *food*, *chicken* e *sandwich* também aparecem com frequência em *tips* positivas das duas bases, o que demonstra a preferência dos usuários por esses locais.

(a) Base manualmente rotulada (b) Base com emoticons.

Figure 3: Palavras mais utilizadas em *tips* positivas

(a) Base manualmente rotulada (b) Base com emoticons.

Figure 4: Palavras mais utilizadas em *tips* negativas

Foram analisados também os principais assuntos ou tópicos discutidos pelas *tips* de cada uma das bases, considerando também a categoria do *venue* ao qual a *tip* foi associada[8]. Para isso, foi utilizado o léxico LIWC - *Linguistic Inquiry and Word Count* [21]. O LIWC é uma ferramenta para análise de texto que avalia componentes emocionais, cognitivos e estruturais de um texto através do uso de um dicionário de mais de 2.300 palavras classificadas em 80 tópicos ou categorias. Esse léxico vem sendo muito explorado em vários contextos principalmente para análise de sentimento em redes sociais [17, 23]. Para essa análise foram considerados 12 desses tópicos que descrevem aspectos psicológicos (emotivos, perceptivos e processos biológicos) e de interesse pessoal (trabalho, lazer, casa e família), por estarem relacionados ao objetivo de classificar as *tips* em positivas e negativas. Foram geradas as distribuições de tópicos para cada uma das 4 categorias de *venues* mais frequentes, utilizando as frequências relativas de cada tópico do LIWC nas palavras das *tips*. As Figuras 5 e 6 mostram essas distribuições separadas por polaridade de *tips* nas duas bases. Para cada curva representando uma categoria de *venue*, um valor mais alto (mais externo) em um dos eixos implica em maior frequência do tópico correspondente nas *tips* associadas a *venues* daquela categoria.

Os resultados indicam que dinheiro (*money*) e trabalho (*work*) são assuntos muito frequentes tanto em *tips* positivas quanto em negativas, nas duas bases. Em particular, dinheiro é muito proeminente na categoria associada a lojas e serviços (*Shops & Services*) enquanto assuntos ligados a trabalho são mais discutidos dentro da categoria relacionada a escritórios (*Professional & Other Places*), com exceção das *tips* negativas da base manual, onde trabalho é mais discutido dentro de locais relacionados a restaurantes (*Food*) e lojas (*Shops & Services*). Pode-se observar uma diferença significativa entre a distribuição dos assuntos discutidos nas *tips* negativas em cada uma das bases. Na base manual, termos relacionados a raiva (*anger*) e a saúde (*health*) são muito utilizados em *tips* das categorias relacionadas a restaurantes e aos escritórios, respectivamente. Já para a base de emoticons, termos ligados a casa (*home*), raiva e tristeza são muito usados em *tips* das categorias que englobam bares e boates (*Nightlife Spots*) enquanto raiva e tristeza são tópicos também frequentes em *tips* negativas ligadas a restaurantes.

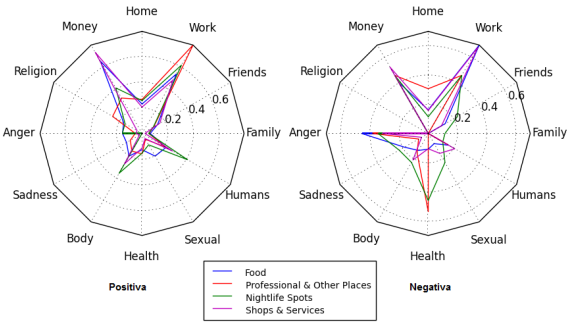


Figure 5: Principais tópicos discutidos nas *tips* positivas (à esquerda) e negativas (à direita) da base manual

Considerando que o foco das duas bases é em *tips* da língua inglesa, foi analisada também a distribuição das polaridade das *tips* do país de língua inglesa com o maior número de tips coletadas (Estados Unidos). As Figuras 7 e 8 mostram as distribuições das *tips* positivas e negativas, respectivamente, por estado americano para a base manualmente rotulada. Os números dentro de cada estado

[8]Para esta análise, foram selecionadas as categorias de *venues* com mais *tips*.

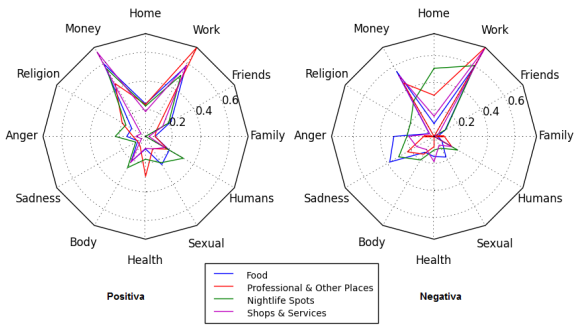


Figure 6: Principais tópicos discutidos nas *tips* positivas (à esquerda) e negativas (à direita) da base de emoticons

no mapa, indicam o número de *tips* postadas em *venues* naquele estado. Califórnia e o de Nova Iorque são os estados com maior concentração de *tips* (positivas e negativas). Dentre os estados com mais *tips* coletadas, Nova Iorque e a Pensilvânia aparecem com a maior proporção de *tips* positivas para negativas (4 positivas para 1 negativa). A conclusão principal destas figuras é que, no geral, há uma tendência muito maior das pessoas adicionarem *tips* positivas, e isto é generalizado por todos os estados do país.

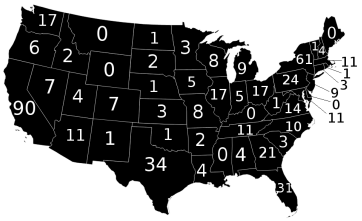


Figure 7: Distribuição de *tips* positivas da base manualmente rotulada por estado americano.

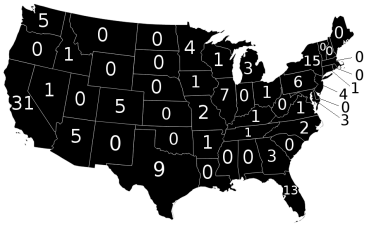


Figure 8: Distribuição de *tips* negativas da base manualmente rotulada por estado americano.

5. DETECÇÃO DE POLARIDADE

Nesta seção são discutidas as técnicas usadas para detecção automática da polaridade (positiva ou negativa) de uma *tip*. Primeiramente, são apresentadas as técnicas de aprendizagem supervisionada. Em seguida, é descrita a técnica não supervisionada baseada em léxico adotada neste trabalho.

5.1 Técnicas Supervisionadas

No geral, as técnicas supervisionadas têm o mesmo funcionamento. Dado um conjunto de treino formado por instâncias (*tips*) que são representadas por vários atributos e previamente rotuladas nas classes de interesse (polaridades), o algoritmo *aprende* um modelo de classificação que pode ser posteriormente aplicado a dados não rotulados (conjunto de teste). Foram analisados 3 algoritmos de classificação automática supervisionada: Naïve Bayes (NB), Máxima Entropia (ME) e Máquina de Vetores Suporte (SVM) [11, 15, 24]. Os três algoritmos são considerados o estado-da-arte em classificação textual.

Para os três algoritmos, cada *tip* foi modelada como uma *bag-of-words*, como em [19]. Entretanto, ao invés de considerar os unigramas da *tip* como atributos, foi utilizada a mesma representação proposta em [1]: cada *tip* t é modelada como um vetor $p_1, ...p_n$, onde p_i é a frequência ponderada da palavra i na *tip* t normalizada pela frequência da palavra i na base de treino (TF-IDF). Experimentos preliminares demonstraram que esta representação leva a resultados superiores aos obtidos usando unigramas ou bigramas. Note que foram removidas as *stopwords* de cada *tip* antes da representação. Os valores de p_i são os atributos explorados pelos algoritmos de classificação.

O Naïve Bayes (NB) é um classificador probabilístico, baseado na aplicação do teorema de Bayes, que tenta inferir as probabilidades de um novo documento (ou *tip*) pertencer a cada uma das classes (polaridades) definidas [24]. Ele é utilizado em diversas aplicações como, por exemplo, filtragem de spam, diagnósticos de doenças e para classificação da polaridade de textos [3, 19]. Para este trabalho, foi utilizada a versão multinomial (MBN) desse classificador, que é mais adequada para a classificação de textos [13]. Nessa versão, a probabilidade de uma classe é parametrizada por uma distribuição Multinomial.

A principal desvantagem do Naïve Bayes é a premissa de independência entre os atributos explorados pelo classificador, que é dificilmente observada na prática. O método Máxima Entropia [15] não assume independência entre os atributos e estima as probabilidades fazendo o mínimo de restrições possíveis. As restrições expressam algum tipo de relacionamento entre os atributos e as classes e são derivadas do conjunto de treino. A distribuição de probabilidade que melhor satisfaz as restrições é aquela com a maior entropia.

Finalmente, o modelos baseados em Máquina de Vetores Suporte (SVM) [11], que também são muito utilizados para classificação em textos, tentam encontrar o melhor hiperplano, definido no espaço dos atributos, que separa as instâncias do treino e que tem a maior distância (margem) entre essas instâncias. O SVM permite a utilização de várias funções de *kernel*, que auxiliam a resolução de vários tipos de problemas. Nós utilizamos aqui o *kernel* linear, já que o número de instâncias é menor que o número de atributos.

Foram utilizados nos experimentos as versões do Naïve Bayes e Máxima Entropia implementadas pela ferramenta *scikit-learn*[9] e a versão do SVM disponível no pacote LIBSVM[10].

5.2 Técnica Não Supervisionada

Os métodos supervisionados precisam de instâncias pré-rotuladas (treino) para o desenvolvimento dos classificadores. Os modelos não supervisionados minimizam essa necessidade, pois exploram o conteúdo do texto para realizar a classificação. Nessa seção é descrita a abordagem não supervisionada baseada em léxico que foi utilizada em nossos experimentos.

[9] http://scikit-learn.org/
[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Léxicos de opiniões são listas de palavras que expressam algum tipo de opinião positiva ou negativa (e.g., "amazing" ou "bad"). Essas listas são muito utilizadas por métodos de detecção de polaridade [16], que tipicamente definem a polaridade pelo número de palavras positivas ou negativas. Os léxicos são construídos manualmente para cada contexto de aplicação. Porém, diferentemente dos conjuntos de treino explorados por técnicas supervisionadas, que são tipicamente específicos da aplicação alvo, léxicos podem ser genéricos o bastante para serem úteis para estudos de aplicações e contextos diferentes. Logo, o custo de construção do léxico pode ser amortizado em um número muito maior de aplicações e investigações. Um aspecto a se considerar na escolha de um léxico é o número de termos incluídos nele ou sua cobertura. Este número pode impactar a eficácia dos métodos que fazem uso do léxico [16]. Assim, vários trabalhos [3, 7, 16, 23] que utilizam métodos não supervisionados para detecção de polaridade fazem uso de léxicos já conhecidos como o SentiWordNet [4] e o LIWC [22]. Para o nosso trabalho, foi escolhido o SentiWordNet como base da técnica não supervisionada avaliada pois sua cobertura é maior que a do LIWC.

O SentiWordNet é um recurso léxico para mineração de opinião baseado em outra ferramenta léxica, o WordNet [14]. WordNet é um dicionário léxico em inglês que agrupa nomes, verbos, adjetivos e advérbios em conjuntos de sinônimos, ou *synsets*, cada qual expressando um conceito distinto[11]. O SentiWordNet associa três pontuações - positivo, negativo e neutro - para cada *synset* do WordNet, representando a força de um sentimento positivo, negativo ou neutro associado àquela palavra. Essas pontuações estão normalizados entre 0 e 1, de tal forma que a soma seja 1.

A abordagem não supervisionada consiste dos seguintes passos para classificar uma dada *tip* em positiva ou negativa:

1. Classificação Gramatical e Lematização[12].: cada palavra de uma *tip* é associada a uma única classe gramatical, tal como adjetivo, advérbio, verbo e pronome através da utilização de *parser part-of-speech*. Em seguida, cada palavra da *tip* é convertida à sua forma canônica, por exemplo, verbos no passado vão para sua forma infinitiva.
2. Tratamento de negação:[12] a polaridade de uma palavra pode ser influenciada caso um termo de negação (por exemplo, *not*) a anteceda. Para lidar com esse cenário, é construída uma árvore de dependência que modela as relações gramaticais de cada palavra ou frase da *tip*. Assim é possível identificar quais palavras da *tip* são influenciadas pelo termo de negação. Essas palavras têm as pontuações positiva e negativa, fornecidas pelo SentiWordNet, invertidas.
3. Sentido da palavra: uma palavra no *SentiWordNet* pode possuir múltiplos *synsets* associadas à mesma classe gramatical. Para contornar esse problema, foi considerada a média das pontuações de todos os *synsets* associados àquela palavra.
4. Polaridade da *tip*: atribui-se pontuações positiva, negativa e neutra à *tip*, cada uma calculada como a média das pontuações correspondentes dos *synsets* de todas as palavras da *tip* que foram encontradas no SentiWordNet. Uma *tip* é considerada com polaridade positiva se a sua pontuação final positiva é maior que pontuação negativa, caso contrário ela é considerada negativa e vice-versa. Se os valores das médias da pontuação positiva e negativa empatarem ou se houver apenas pontuação neutra, a polaridade da *tip* é considerada indefinida e essas *tips* são descartadas.

[11]Cada possível significado de uma mesma palavra está em *synsets* diferentes

[12] Esta fase foi implementada usando a ferramenta disponível em `http://www-nlp.stanford.edu/software/corenlp.shtml`.

6. AVALIAÇÃO

Esta seção descreve a metodologia de avaliação adotada na análise dos métodos de detecção de polaridade de *tips* (Seção 6.1) e discute os principais resultados obtidos (Seção 6.2).

6.1 Metodologia de Avaliação

A avaliação dos métodos apresentados na Seção 5 foi feita utilizando validação cruzada de 5 partes em cada uma das bases de *tips* descritas na Seção 3. Em outras palavras, cada base foi dividida em 5 partes, das quais 4 foram usadas como conjunto de treino e a parte restante foi usada como teste. O processo foi repetido 5 vezes, utilizando cada uma das partes como teste e produzindo assim 5 resultados. O conjunto de treino foi utilizado apenas para as abordagens supervisionadas, para "aprender"os modelos de classificação. Todos os métodos foram avaliados sobre os conjuntos de testes. Para evitar o desbalanceamento entre as classes (positiva e negativa), que afeta a acurácia da predição, foi utilizado a técnica de *undersampling* [8], em que a menor classe determina o número de instâncias de cada classe usadas para o treino. Assim, para cada rodada da validação cruzada foi efetuada 5 amostragens de cada uma das classes do treino. Desse modo foram produzidos 25 resultados diferentes para cada base de *tips*. A próxima seção apresenta resultados para cada abordagem que são valores médios desses 25 experimentos juntamente com intervalos de confiança de 95%.

A acurácia de cada método foi avaliada utilizando três métricas: precisão, revocação e F1 [2]. A precisão p de uma classe c é o número de *tips* corretamente classificadas na classe c sobre o total de *tips* preditas como sendo da classe c. A revocação r de uma classe c é o número de *tips* corretamente classificadas na classe c sobre o número de *tips* na classe c. A métrica F1 é a média harmônica, $2pr/(p+r)$, entre a precisão p e revocação r, sintetizando o valor das duas métricas. Foram computados valores de precisão, revocação e F1 para cada classe (polaridade) separadamente, assim como valores médios para as duas classes.

6.2 Resultados

As Tabelas 2 e 3 mostram os resultados de cada uma das abordagens supervisionadas, Naïve Bayes (NB), Máxima Entropia (ME) e SVM, e não supervisionada (Léxico) para as bases de *tips* manualmente rotuladas e com emoticons, respectivamente. Os melhores resultados (incluindo empates estatísticos) são mostrados em negrito. A significância desses valores foi testada utilizando um teste-t pareado [10] considerando uma confiança de 95%.

Considerando a precisão por classe de polaridade, pode-se observar que para ambas as bases, os melhores valores são aqueles obtidos pelos métodos supervisionados Naïve Bayes e SVM. Em particular, o método Naïve Bayes apresenta ganhos de até 3,51% e 2,78% sobre os demais métodos em *tips* da base manual e na base de emoticons, respectivamente, e não apresenta diferença estatisticamente significativa com o método SVM em ambas as bases. Para a classe negativa, os melhores valores de precisão na base manual ocorrem quando a abordagem não supervisionada baseada em léxico é utilizada (ganhos de até 22,68% sobre os demais métodos), enquanto que para a base de emoticons, o léxico se apresenta empatado com o SVM e o Naïve Bayes. Note que, no geral, os valores de precisão para a classe negativa são menores que os da classe positiva. Isto ocorre devido ao desbalanceamento entre as classes (Tabela 1), principalmente na base de emoticons. Este desbalanceamento leva a uma dominância da maior classe (positiva) sobre os resultados da classificação. Finalmente, considerando a precisão média não foi observada diferença estatisticamente significativa entre o Naïve Bayes (supervisionado) e o método baseado no léxico (não supervisionado) para a base manualmente rotulada. Já para

Table 2: Resultados da base de *tips* manualmente rotuladas

	Método	Classe Positiva	Classe Negativa	Média
Precisão	Naïve Bayes	**0,9173±0,0067**	0,4333±0,0180	**0,6753±0,0103**
	Máxima Entropia	0,9017±0,0092	0,3950±0,0206	0,6484±0,0120
	SVM	**0,9097±0,0096**	0,4169±0,0221	0,6633±0,0127
	Léxico (SentiWordNet)	0,8861±0,0121	**0,4846±0,0390**	**0,6853±0,0189**
Revocação	Naïve Bayes	0,7311±0,0126	**0,7547±0,0241**	**0,7429±0,0119**
	Máxima Entropia	0,7015±0,0201	0,7124±0,0364	0,7070±0,0146
	SVM	0,7176±0,0270	**0,7302±0,0387**	0,7239±0,0157
	Léxico (SentiWordNet)	**0,8183±0,0180**	0,6159±0,0276	0,7171±0,0154
F1	Naïve Bayes	0,8133±0,0083	**0,5496±0,0190**	**0,6814±0,0116**
	Máxima Entropia	0,7879±0,0116	0,5058±0,0222	0,6469±0,0135
	SVM	0,8003±0,0166	**0,5278±0,0226**	0,6640±0,0157
	Léxico (SentiWordNet)	**0,8502±0,0118**	**0,5369±0,0313**	**0,6935±0,0187**

a base de emoticons, os métodos supervisionados SVM e Naïve Bayes são os que produzem os melhores resultados com ganhos (estatisticamente significativos) de até 3,38%.

Em termos de revocação para a classe positiva, o método baseado em léxico supera os métodos supervisionados em até 16,65% para a base manual e em até 14,71% para a base de emoticons. No entanto, para a classe negativa, a melhor revocação ocorre com os métodos supervisionados Naïve Bayes e SVM (empate estatístico) em ambas as bases, com ganhos de até 37,78%. Os grandes ganhos na classe negativa levam a uma superioridade dos métodos supervisionados considerando a revocação média nas duas classes: os melhores resultados na base manual foram obtidos com o Naïve Bayes, enquanto que, na base de emoticons, Naïve Bayes e SVM aparecem empatados como os melhores métodos.

Considerando a métrica F1, que combina revocação e precisão, o método baseado em léxico produz os melhores resultados para *tips* da classe positiva em ambas as bases. Para a classe negativa, foi observado um empate entre Naïve Bayes, SVM e o método baseado em léxico na base manual, enquanto que na base de emoticons esse empate é somente entre Naïve Bayes e SVM. No geral, Naïve Bayes e o método baseado em léxico aparecem empatados como os melhores métodos em termos de F1 médio nas duas bases, enquanto que na base de emoticons este empate também inclui o SVM.

Os resultados obtidos podem ser sumarizados como segue:

- O método não supervisionado baseado em léxico apresenta resultados estatisticamente superiores ou empatados, em termos do F1 médio, com os melhores métodos supervisionados (Naïve Bayes e SVM) nas duas bases de *tips*. Assim, se o objetivo da aplicação é recuperar tanto *tips* positivas quanto *tips* negativas, a um custo baixo (sem rotulação manual), o método baseado em léxico é preferível.
- O método baseado em léxico melhora o F1 da classe positiva em até 7,91%. Logo, esse método deve ser utilizado para aplicações cujo foco é recuperação de *tips positivas*.
- Se o foco é a detecção de *tips negativas*, os melhores métodos são os supervisionados, em especial, os métodos Naïve Bayes e SVM. Para a base com emoticons, esses métodos supervisionados apresentam um ganho de até 11,11% na detecção de *tips* negativas.
- Todos os métodos produzem bons resultados para as três métricas na base rotulada manualmente (diferença de até 19,5%), o que pode refletir um maior ruído (p.ex: sarcasmo) e uma maior incerteza na rotulação automática via emoticons que afeta a eficácia dos métodos.

Em suma, neste trabalho foi mostrado que no contexto de microrevisões, mais especificamente *tips* no Foursquare, o método não supervisionado obteve, no geral, uma acurácia comparável à dos melhores métodos supervisionados (Naïve Bayes e SVM), sem os custos associados a esses últimos. Entretanto, é preciso ressaltar que a escolha do melhor método deve levar em consideração os custos e restrições de cada tipo de abordagem. Um método supervisionado geralmente requer que um conjunto de treino manualmente rotulado seja construído. A rotulação automática, explorando por exemplo emoticons, pode ser feita, mas sua eficácia está limitada à presença de emoticons na *tip*[13] assim como sujeita a ruído e maior incerteza. Já o método não supervisionado analisado requer a disponibilidade de um léxico para o idioma alvo. Além disso, para este método, existe a restrição do tamanho do vocabulário que é coberto pelo léxico, o que pode fazer com que certas *tips* não possam ser classificadas. Em particular, o método de detecção de polaridade baseado no léxico SentiWordNet, publicamente disponível [16] e amplamente usado na literatura, não pode ser aplicado em 1,4% e 1,6% das *tips* originalmente obtidas nas bases manual e de emoticons respectivamente, já que nenhuma das palavras usadas nestas *tips* estavam presentes no léxico[14].

7. CONCLUSÃO

Neste trabalho foram analisados métodos para detecção de polaridade ou sentimento de *tips* do Foursquare. Na literatura sobre análise de sentimento não existe um consenso sobre qual é a melhor abordagem para textos curtos, como as *tips* no Foursquare. Logo, foram avaliados tanto métodos supervisionados - os algoritmos Naïve Bayes, Máxima Entropia e SVM - quanto um método não supervisionado baseado no léxico SentiWordNet.

Estes métodos foram avaliados em duas bases de *tips*: uma rotulada manualmente e outra rotulada automaticamente explorando a presença de emoticons. Os resultados da avaliação mostraram que, no geral, o método não supervisionado baseado em léxico obteve resultados comparáveis aos produzidos pelos melhores métodos supervisionados (Naïve Bayes e SVM) sem os custos e limitações de construção do conjunto de treino associados a esses últimos. No entanto, a escolha do método deve levar em consideração objetivos específicos: se o objetivo é detectar principalmente *tips* positivas, o método não supervisionado é o mais indicado, enquanto os métodos supervisionados (Naïve Bayes e SVM) são os mais indicados para a detecção de *tips* negativas. Esta escolha deve levar em conta

[13]Somente 3,39% das *tips* de língua inglesa na nossa base de 10 milhões de *tips* contêm emoticons.

[14]Estas *tips* foram filtradas das duas bases (vide Seção 3).

Table 3: Resultados da base de *tips* rotuladas com emoticons

	Método	Classe Positiva	Classe Negativa	Média
Precisão	Naïve Bayes	**0,9393±0,0033**	**0,2457±0,0110**	**0,5925±0,0056**
	Máxima Entropia	0,9270±0,0040	0,2193±0,0098	0,5731±0,0050
	SVM	**0,9399±0,0040**	**0,2394±0,0133**	**0,5896±0,0065**
	Léxico (SentiWordNet)	0,9139±0,0054	**0,2416±0,0127**	0,5778±0,0070
Revocação	Naïve Bayes	0,6888±0,0078	**0,6936±0,0181**	**0,6912±0,0087**
	Máxima Entropia	0,6670±0,0078	0,6400±0,0156	0,6535±0,0085
	SVM	0,6698±0,0227	**0,7028±0,0266**	**0,6863±0,0096**
	Léxico (SentiWordNet)	**0,7651±0,0055**	0,5101±0,0247	0,6376±0,0122
F1	Naïve Bayes	0,7946±0,0051	**0,3623±0,0133**	**0,5785±0,0078**
	Máxima Entropia	0,7756±0,0055	0,3261±0,0118	0,5508±0,0075
	SVM	0,7807±0,0153	**0,3554±0,0150**	**0,5681±0,0128**
	Léxico (SentiWordNet)	**0,8328±0,0040**	0,3271±0,0156	**0,5800±0,0086**

também a disponibilidade de recursos para o uso de cada método. Em particular os custos de construção manual ou as limitações da construção automatizada (p.ex: explorando emoticons) de um conjunto de treino precisam ser considerados para a adoção de um método supervisionado, enquanto a disponibilidade e a cobertura de um léxico para a língua alvo têm que ser avaliados para a escolha do método não supervisionado.

Como trabalhos futuros, pretendemos avaliar os métodos em outras bases de micro-revisões para validação dos resultados e desenvolver uma solução híbrida que combine as abordagens supervisionada e não supervisionada.

8. AGRADECIMENTOS

Esta pesquisa é parcialmente financiada pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb (MCT/CNPq 573871/2008-6), CNPq, CAPES e FAPEMIG.

9. REFERENCES

[1] F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou. Content vs. Context for Sentiment Analysis: a Comparative Analysis over Microblogs. In *Proc. ACM HT*, 2012.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval - The Concepts and Technology Behind Search*. Pearson Education Ltd., 2011.

[3] A. Bermingham and A. Smeaton. Classifying Sentiment in Microblogs: is Brevity an Advantage? In *Proc. CIKM*, 2010.

[4] A. Esuli and F. Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proc. LREC*, 2006.

[5] A. Go, R. Bhayani, and L. Huang. Twitter Sentiment Classification using Distant Supervision. Technical report, Department of Computer Science, Stanford University, 2009.

[6] P. Guerra, A. Veloso, W. Meira, Jr, and V. Almeida. From Bias to Opinion: a Transfer-Learning Approach to Real-Time Sentiment Analysis. In *Proc. SIGKDD*, 2011.

[7] A. Hamouda and M. Rohaim. Reviews Classification Using SentiWordNet Lexicon. *The Online Journal on Computer Science and Information Technology*, 2(4):120–123, 2011.

[8] H. He and E. A. Garcia. Learning From Imbalanced Data. *IEEE Trans. on Knowledge and Data Engin.*, 21(9):1263–1284, 2009.

[9] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised Sentiment Analysis with Emotional Signals. In *Proc. WWW*, 2013.

[10] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1991.

[11] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features Machine Learning. In *Proc. ECML*, 1998.

[12] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic Construction of a Context-Aware Sentiment Lexicon: an Optimization Approach. In *Proc. WWW*, 2011.

[13] A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *Proc. AAAI/ICML Workshop Learning for Text Categorization*, 1998.

[14] G. A. Miller. WordNet: a Lexical Database for English. *Communications of ACM*, 38(11):39–41, 1995.

[15] K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification. In *Proc. IJCAI*, 1999.

[16] B. Ohana and B. Tierney. Sentiment Classification of Reviews using SentiWordNet. In *Proc. IT & T*, 2009.

[17] G. Paltoglou and M. Thelwall. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. *ACM Trans. on Intelligent Systems and Technology*, 3(4):66:1–66:19, 2012.

[18] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[19] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proc. EMNLP*, 2002.

[20] J. Read. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proc. ACL Student Research Workshop*, 2005.

[21] Y. Tausczik and J. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[22] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 2010.

[23] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting Elections with Twitter: What 140 characters Reveal about Political Sentiment. In *Proc. ICWSM*, 2010.

[24] H. Zhang. Exploring Conditions For The Optimality Of Naïve Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(2):183–198, 2005.