

# Métodos para Análise de Sentimentos no Twitter

Matheus Araújo  
UFMG  
Belo Horizonte, Brasil  
matheus.araujo@dcc.ufmg.br

Pollyanna Gonçalves  
UFMG  
Belo Horizonte, Brasil  
pollyannaog@dcc.ufmg.br

Fabrcio Benevenuto  
UFMG  
Belo Horizonte, Brasil  
fabrcio@dcc.ufmg.br

## ABSTRACT

Sentiment analysis has been used in several applications including the analysis of the repercussion of events in online social networks (OSNs), as well as to summarize public perception about products and brands on discussions on those systems. There are multiple methods to measure sentiments, varying from lexical-based approaches to machine learning methods. Despite the wide use and popularity of some of those methods, it is unclear which method is better for identifying the polarity (i.e. positive or negative) of a message, as the current literature does not provide a comparison among existing methods. This comparison is crucial to allow us to understand the potential limitations, advantages, and disadvantages of popular methods in the context of OSNs messages. This work aims at filling this gap by presenting a comparison between 8 popular sentiment analysis methods. Our analysis compares these methods in terms of coverage and in terms of correct sentiment identification. We also develop a new method that combines existing approaches in order to provide the best coverage results with competitive accuracy. Finally, we present iFeel, a Web service which provides an open API for accessing and comparing results across different sentiment methods for a given text.

## Categories and Subject Descriptors

J.4. [Computer Applications]: Social and behavioral sciences Miscellaneous; H.3.5 [Online Information Services]: Web-based services

## General Terms

Human Factors, Measurement.

## Keywords

Twitter, emoticons, análise de sentimentos, redes sociais.

## 1. INTRODUÇÃO

Redes sociais online têm se tornado uma importante plataforma de comunicação que agrupa diversas informações, entre elas opiniões e sentimentos expressos por seus usuários em simples conversas

ou mensagens. A quantidade de usuários ativos e o volume de dados criados diariamente nessas redes é impressionante. Uma plataforma popular nos dias de hoje é o Twitter que, sozinho, possui mais de 200 milhões de usuários, que compartilham cerca de 400 milhões de tweets<sup>1</sup> por dia [25]. Nesse contexto, pesquisadores e empresas conseguem coletar esses dados para análises de conteúdo em grande escala [11].

Diversos estudos no contexto de redes sociais estão focados na identificação e monitoramento de polaridade em mensagens compartilhadas, partindo da hipótese que a quantidade expressiva de dados postados numa parcela significativa estaria relacionada ao humor e a emoções expressas pelos usuários. Análise de polaridade em mensagens possui inúmeras aplicações, especialmente no desenvolvimento de sistemas capazes de capturar opiniões públicas relacionadas a eventos sociais [16] e até mesmo lançamentos de produtos em tempo real.

Entretanto, pouco se sabe sobre como os vários métodos propostos funcionam no contexto das redes sociais online. Métodos para análise de sentimentos vem sendo muito utilizados para desenvolver aplicações sem estudo prévio a respeito da aplicabilidade do método no contexto desejado, assim como suas vantagens, desvantagens e potenciais limitações quando comparados a outros métodos. De fato, muitos desses métodos foram propostos para análise de sentenças longas e não para análise de mensagens curtas em tempo real. Além do mais, poucos esforços foram feitos com o objetivo de comparar tais métodos.

Nesse trabalho, temos como objetivo preencher essa lacuna na comparação de métodos para análise de sentimento. Utilizamos 2 bases de dados diferentes provenientes de redes sociais online para comparar 8 métodos propostos na literatura: LIWC, Happiness Index, SentiWordNet, SASA, PANAS-t, Emoticons, SenticNet e SentiStrength. A primeira base consiste de cerca de 1,8 bilhões de mensagens coletadas do Twitter [11], representando um histórico completo da rede no período coletado. Dessa base de dados fomos capazes de filtrar tweets associados a 6 eventos sociais relacionados a tragédias, lançamento de produtos, política, saúde e esporte. A segunda base de dados consiste de uma coleção de textos rotulados por humanos para positivo e negativo [22]. A partir de bases de dados reais, comparamos os 8 métodos para análise de sentimentos em termos de abrangência (a fração de mensagens capturadas por cada método) e concordância (a fração de sentimentos corretamente identificados por cada método).

Entre os vários resultados encontrados, podemos sumarizar alguns deles:

<sup>1</sup>Mensagens com no máximo 140 caracteres compartilhadas na rede social online Twitter.

**Tabela 1: Emoticons e suas variações**

- Os métodos possuem diferentes graus de abrangência, variando entre 4% e 95% quando aplicados a dados associados a eventos reais. Isso sugere que, dependendo do método utilizado, apenas uma pequena fração de mensagens será analisada, podendo levar a resultados enviesados ou não representativos.
- Nenhum método alcançou níveis altos de abrangência e concordância ao mesmo tempo. O método Emoticons atingiu a maior acurácia (acima de 85%), porém uma das menores abrangências (4–13%).
- A concordância dos métodos, quando aplicados aos dados rotulados, variaram entre 33% e 80%, sugerindo que uma mesma amostra de dados pode ser interpretada de forma diferente dependendo do método escolhido.
- Existe desacordo entre os métodos na predição de sentimentos para diferentes eventos considerados. Para o caso do evento da queda de um avião, metade dos métodos detectaram mais positividade do que negatividade. O mesmo é observado em outros eventos onde eram esperados uma maior quantidade de sentimentos negativos.

Baseados nessas observações, desenvolvemos um novo método para análise de sentimentos que consiste da combinação dos métodos estudados com objetivo de alcançar maior abrangências e acurácia competitivas com relação aos métodos existentes. O restante do artigo está organizado como se segue. A seguir, descrevemos os 8 métodos utilizados para comparações. Depois, apresentamos as métricas para comparações, assim como uma descrição sobre as bases de dados utilizadas. Na seção 4 apresentamos os resultados alcançados no processo de comparação, e então propomos um método para detecção de sentimentos. Por fim, concluímos e apresentamos direções para trabalhos futuros.

## 2. MÉTODOS




Esta seção apresenta uma breve descrição dos 8 métodos para análise de sentimentos que são discutidos neste trabalho.

### 2.1 Emoticons

Talvez o jeito mais simples de identificar polaridade de uma mensagem seja baseado na análise de emoticons [14]. Nos últimos anos, emoticons tem se tornado tão populares que alguns foram adicionados ao conhecido Dicionário de Oxford [2]. Emoticons são principalmente baseados em faces e podem expressar sentimentos de felicidade ou tristeza, embora uma grande quantidade deles não representam faces, por exemplo o emoticon <3 que representa um coração, que expressa amor ou afeição.

Para extrair polaridade de emoticons utilizamos um conjunto dos emoticons populares em sites como Yahoo e MSN [4, 1], descritos na Tabela 1. A tabela também inclui as variações mais comuns para emoticons que expressam polaridade positiva, negativa e neutra. Mensagens com mais de um emoticon foram associadas à polaridade do primeiro emoticon encontrado. No entanto esses casos ocorreram poucas vezes e foram irrelevantes na nossa base de dados.

Como podemos esperar, a taxa de mensagens em redes sociais online que contém pelo menos 1 emoticon é muito baixa se comparado ao total de mensagens que poderiam expressar algum sentimento. Trabalhos recentes verificaram que essa taxa é menor que

Emoticon	Polaridade	Símbolos
	Positivo	:) :D :o) :o] :o] :o] :-] :-] :-] =) =) =^] =^] =^] :B :-D :-B :^D :^B :^D :^B =B =^B =^D :') :') :') =') =') =') <3 ^.^ ^.^ ^.^ ^.^ :* =* :-* ;) ;) ;) :-p :-p :-b :-b :^p :^p :^b =^p =^p \o\ /o/ :p :p :b =b =^p =^p \o/
	Negativo	D: D= D-: D^: D^= : ( : [ : [ :o ( :o [ :^ ( :^ [ :^ [ =^ ( =^ [ >= ( >= [ >= { >= ( >:- { >:- [ >:- ( >=^ [ >:- ( :- [ :- ( = ( = [ = [ =^ [ >:-= ( >= [ :/ ( :/ [ :/ { =/ ( =/ ( =/ [ =\ : \ =/ :/ =\\$ o.o O_o O_o :\$:- { >:- { >=^ ( >=^ [ :o {
	Neutro	:  =  :-  >.< >.> >.< :o :o =o :@ =@ :^o :^@ -.- -.-' -.- -.-' :x =X =# :-x :-@ :-# :^x :^# :#

10% [18, 14]. Entretanto, emoticons vem sendo frequentemente utilizados em combinações com outros métodos para a construção de bases de treinamento para técnicas de aprendizado de máquina supervisionada [19].

### 2.2 LIWC

LIWC (Linguistic Inquiry and Word Count) [21] é uma ferramenta para análise de texto que estima componentes emocionais, cognitivos e estruturais de um dado texto baseada no uso de dicionários contendo palavras e suas respectivas categorias. A título de exemplo, no LIWC a palavra “agree” pertence a 5 categorias: *assent*, *afffective*, *positive emotion*, *positive feeling*, e *cognitive process*. Portanto, além de detectar *positive* ou *negative feeling* em um texto, o LIWC também fornece outras categorias de palavras.

A ferramenta é unicamente comercial e fornece funções otimizadas, como a permissão para inclusão de dicionários personalizados. Para este trabalho, utilizamos a versão LIWC2007, a versão mais recente do sistema, e o dicionário padrão para o idioma Inglês, que consiste de 4.500 palavras e mais de 100 categorias. O software pode ser encontrado em <http://www.liwc.net/>. Com o objetivo de medir polaridade, examinamos a taxa de sentimentos positivos e negativos em termos das categorias de *emotion* e *affective*.

### 2.3 SentiStrength

Entre os métodos baseados em abordagens de aprendizado de máquina está o SentiStrength [22], que compara métodos de classificação supervisionadas e não-supervisionadas. O trabalho utilizou para classificação uma versão expandida do dicionário do LIWC [21], com a adição de novas características para o contexto de redes sociais. Essas características incluem conjuntos extras de palavras positivas e negativas, um conjunto de palavras que dão maior entonação para um sentimento (ex.: “very” ou “somewhat”), um conjunto de emoticons com polaridade associadas e percepções de pontuações repetidas (ex.: “Cool!!!!”). Para avaliar o método, autores utilizaram bases de dados rotuladas de 6 diferentes fontes Web 2.0: MySpace, Twitter, posts no Digg, comentários no fórum da BBC e Runners Word, e Youtube.

O SentiStrength consiste da combinação das técnicas que produziram melhor resultados entre as citadas acima. Neste trabalho utilizamos a versão 2.0 do método, que está disponível em [3].

### 2.4 SentiWordNet

SentiWordNet [13] é uma ferramenta muito utilizada em mineração de opinião, e é baseado no dicionário léxico WordNet [17]. Esse dicionário agrupa adjetivos, verbos e outras classes gramaticais em conjuntos chamados *synset*. O SentiWordNet associa a cada *synset* do WordNet três valores de pontuação que indicam o sentimento de

um texto: positivo, negativo e objetivo (neutralidade). Cada pontuação é obtida utilizando um método de aprendizagem de máquina semi-supervisionada, e variam de 0 a 1, com soma igual a 1. Para melhor entender o funcionamento do método, suponha que para um dado *synset*  $s = [bad, wicked, terrible]$  tenha sido extraído de um tweet. O resultado obtido pelo método é 0,000 para positividade, 0,850 para negatividade e 0,150 para objetividade, respectivamente. A avaliação do SentiWordNet foi feita utilizando-se um dicionário léxico rotulado.

Nesse artigo utilizamos a versão 3.0 do SentiWordNet, disponível em <http://sentiwordnet.isti.cnr.it/>. Para associar polaridade baseados nesse método, consideramos a média da pontuação dos *synsets* e diremos que um texto dado é positivo se o valor resultante para positivo for maior que o valor encontrado para negativo. Pontuações para objetividade não foram consideradas nesse trabalho para determinar polaridade.

## 2.5 SenticNet

SenticNet [10] é um método para mineração de opinião e análise de sentimentos que explora técnicas de Inteligência Artificial e Web Semântica. O objetivo do SenticNet é inferir polaridade de textos em nível semântico, e não sintático. O método utiliza técnicas de Processamento de Linguagem Natural (PLN) para criar significados semânticos ou polaridade para aproximadamente 14.000 conceitos, nome dado pelos autores. Por exemplo, para interpretar a mensagem “Boring, it’s Monday morning”, SenticNet primeiramente tenta identificar conceitos, que neste caso seriam “boring” e “Monday morning”. E então calcula a polaridade para cada conceito, nesse caso -0,383 para “boring”, e +0,228 para “Monday morning”. O resultado final para sentimentos no dado exemplo seria de -0,077, que consiste na média dos valores encontrados para cada conceito.

O SenticNet foi testado pelos autores como uma ferramenta para medição de níveis de polaridade em opiniões de pacientes sobre o *National Health Service* na Inglaterra [9]. Autores também testaram o método em base de dados coletadas do LiveJournal, onde mensagens foram rotuladas em 130 estados de humor pelos próprios usuários, e que foram transformados em positivo e negativo [19, 20]. Nosso trabalho utiliza a versão 2.0 do SenticNet, disponível em <http://sentic.net/>.

## 2.6 SASA

Empregamos ao trabalho mais uma técnica baseada em aprendizado de máquina, o SailAil Sentiment Analyzer (SASA) [23]. O SASA foi originalmente proposto como um método para análise de 17.000 tweets rotulados associados as eleições norte-americanas de 2012. A ferramenta, de código aberto, foi avaliada no Amazon Mechanical Turk (AMT) [5], onde *turkers* rotularam tweets como positivos, negativos, neutros ou indefinidos. Esse método foi acrescentado a nossa análise por ser uma ferramenta aberta e ainda não ter sido comparado com nenhum outro método para análise de sentimentos da literatura. Utilizamos o pacote Python SASA na versão 0.1.3, disponível em <https://pypi.python.org/pypi/sasa/0.1.3>.

## 2.7 Happiness Index

Happiness Index [12] consiste de uma escala de sentimentos que utiliza o popular *Affective Norms for English Words* (ANEW) [8]. O ANEW é uma coleção de 1.034 palavras associadas a dimensões afetivas de valência, excitação e dominância. Happiness Index foi construído baseado no ANEW e calcula pontuações com valores entre 1 e 9 para um texto dado, indicando a “quantidade” de felicidade que existe naquele texto. Autores calcularam a frequência

em que cada palavra do ANEW aparece no texto e então computa o peso médio encontrado, levando em consideração apenas o sentimento de valência. Para validação, autores aplicaram o método em letras e títulos de músicas e mensagens de blogs. Como resultados, autores encontraram que níveis de felicidade em letras de músicas tiveram um decréscimo entre 1961 e 2007, mas aumentou nas amostras extraídas de blogs.

Com objetivo de adaptar o Happiness Index para detectar polaridade, consideramos que qualquer texto classificado pelo método no intervalo [1..5) como sendo negativo e [5..9] como sendo positivo.

## 2.8 PANAS-t

O PANAS-t [15] é uma escala psicométrica para detecção de humor que captura flutuações de humor no Twitter. O método consiste de uma versão adaptada do *Positive Affect Negative Affect Scale* (PANAS) [24], que é uma escala bastante conhecida na psicologia. O PANAS-t é baseado em um largo conjunto de palavras associadas a 11 sentimentos: jovialidade, autoconfiança, serenidade, surpresa, medo, tristeza, culpa, hostilidade, timidez, fadiga e atenção. O método foi desenvolvido para detectar qualquer acréscimo ou decréscimo de sentimentos ao longo de um período.

Para associar textos a sentimentos específicos, o PANAS-t primeiramente utiliza uma base de comparações para cada sentimento baseados em uma coleta completa do Twitter. Em seguida, o método calcula a pontuação  $P(s)$  para cada sentimento  $s$  em um dado período, com valores resultantes entre  $[-1, 0; 1, 0]$  para indicar a variação desse sentimento. Por exemplo, dado um conjunto de tweets que contém  $P(“surpresa”) = 0,250$ , isso significaria que o sentimento de surpresa teve um acréscimo de 25% quando comparado a um dia típico. De forma análoga,  $P(s) = -0,015$  significa que houve um decréscimo de 1,5% do sentimento  $s$ . Para avaliação do método, os autores apresentaram evidências do seu bom funcionamento em eventos globais populares. Neste trabalho consideramos os sentimentos de jovialidade, autoconfiança, serenidade e surpresa como sentimentos positivos, e medo, tristeza, culpa, hostilidade, timidez e fadiga como sentimentos negativos. O sentimento de atenção foi considerado neutro, e foi desconsiderado das nossas análises.

Um método similar ao PANAS-t consiste de uma adaptação do *Profile of Mood States* (POMS) [7], uma escala psicológica que mede 6 escalas de humor: tensão, depressão, raiva, vigor, fadiga e confusão. Entretanto, esse método não foi incluído nas nossas análises pois este não está disponível mesmo sob requerimento.

## 3. METODOLOGIA

Após apresentar os 8 métodos para análise de sentimentos que iremos analisar, descrevermos nossa base de dados e as métricas utilizadas para as comparações entre eles.

### 3.1 Base de Dados

Neste trabalho utilizamos 2 bases de dados.

#### 3.1.1 Histórico Completo do Twitter

A primeira base utilizada consiste de um histórico completo de tweets postados no período entre a criação da rede, em 2006 até Agosto de 2009 [11]. Essa base de dados contém cerca 55 milhões de usuários com 1.9 bilhões de links entre eles e quase 1.8 bilhões de tweets postados nesse período. Essa base é apropriada para o nosso propósito pois engloba todos os usuários que configuraram seus tweets como públicos, não consistindo de uma simples amostra, aliviando viés. Mais importante, essa base nos permite analisar a repercussão de eventos populares passados e avaliar os métodos em cenários reais.

**Tabela 2: Sumário de informações dos 6 eventos analisados**

Evento	Período	Palavras-chave
Airfrance	01—06.06.2009	victims, passengers, A330, 447, crash, airplane, airfrance
Eleições-EUA2008	02—06.11.2008	voting, vote, candidate, campaign, mccain, democrat*, republican*, obama, bush
2008Olimpiadas	06—26.08.2008	olympics, medal*, china, beijing, sports, peking, sponsor
Susan Boyle	11—16.04.2009	susan boyle, I dreamed a dream, britain's got talent, les miserables
H1N1	09—26.06.2009	outbreak, virus, influenza, pandemi*, h1n1, swine, world health organization
Harry Potter	13—17.07.2009	harry potter, half-blood prince, rowling

Escolhemos 6 eventos que foram amplamente discutidos por usuários do Twitter<sup>2</sup>. Dentre esses, sumarizados na Tabela 2, há assuntos relacionados a tragédias, estreias, política, saúde e esporte. Para extrair apenas tweets associados a esses eventos identificamos o conjunto de palavras-chave em sites de notícias, blogs, Wikipédia e informações individuais. Dado nossa lista selecionada de palavras-chave, conseguimos filtrar tweets relacionados pela base de dados. Esse processo é similar ao que é aplicado em ferramentas de mineração na coleta de dados associados a tópicos específicos.

Limitamos a duração de cada evento pois palavras-chave populares são tipicamente alvo de *spammers* após certo tempo [6]. A primeira coluna apresenta o nome do evento a que iremos nos referir no restante do artigo. Como a tabela não apresenta um gabarito dos sentimentos associados a cada um dos 6 eventos, iremos utilizar esses dados para comparar sentimentos detectados por todos os métodos.

### 3.1.2 Base Rotulada da Web 2.0

A segunda base de dados contém 6 conjuntos de mensagens rotuladas por humanos como positivas ou negativas, disponibilizada em estudos dos desenvolvedores do método SentiStrength [22]. Essa base de dados inclui uma grande quantidade de textos do MySpace, Twitter, Digg, fórum do BBC e do Runners World, e comentários do Youtube. A Tabela 3 sumariza a quantidade de mensagens em cada base e a fração de sentimentos positivos e negativos rotuladas.

**Tabela 3: Dados rotulados**

Dados	# Mensagens	Pos / Neg
Twitter	4.242	58,58% / 41,42%
MySpace	1.041	84,17% / 15,83%
Youtube	3.407	68,44% / 31,56%
Fórum da BBC	1.000	13,16% / 86,84%
Runners world	1.046	68,65% / 31,35%
Digg	1.077	26,85% / 73,15%

Como essa base de dados rotulada, poderemos analisar a acurácia com que cada método identifica polaridade nesses dados. Em razão do SentiStrength ter sido treinado utilizando esses dados, ele será desconsiderado dessas análises.

## 3.2 Métricas para Comparações

Definimos as métricas utilizadas para avaliar os métodos que estamos analisando considerando os seguintes valores:

		Observação real	
		Positivo	Negativo
Predição esperada	Positivo	a	b
	Negativo	c	d

Sendo *a* o número de mensagens corretamente classificadas como positivas (*true positive*), *b* o número de mensagens negativas classificadas como positivas (*false positive*), *c* o número de mensagens

positivas classificadas como negativas (*false negative*), e *d* o número de mensagens negativas classificadas como negativas (*true negative*). Para comparar e avaliar os métodos, consideraremos as seguintes métricas: taxa de *true positive* (*recall*):  $R = a/(a + c)$ , taxa de falso positivos: (*precision*)  $P = a/(a + b)$ , acurácia:  $A = (a+d)/(a+b+c+d)$ , e F-measure:  $F = 2 \cdot (P \cdot R)/(P + R)$ . Em muitos casos iremos utilizar apenas o F-measure para avaliação, já que essa medida testa acurácia e depende da precisão e *recall*.

Escolhemos utilizar as métricas acima já que elas possuem aplicações diretas. A taxa de *true positive* (*recall*) pode ser entendida como a taxa em que mensagens positivas foram corretamente identificadas (*R*), enquanto que a taxa de *true negative* é entendida como a taxa em que mensagens negativas foram preditas como negativas. A acurácia representa a taxa em que um método identificou sentimentos corretamente (*A*). A taxa de precisão calcula o quão próximo os valores medidos estão um do outro (*P*). Também utilizamos a F-measure para comparar resultados, já que ela relaciona precisão e *recall*. Idealmente, um método para identificação de polaridade alcança o máximo valor para F-measure, 1, significando que a classificação de polaridade foi perfeita.

## 4. RESULTADOS DAS COMPARAÇÕES

Com o objetivo de identificar vantagens, desvantagens e possíveis limitações dos métodos na detecção de polaridade, apresentamos os resultados das comparações feitas sobre eles.

### 4.1 Abrangência

Para cada evento descrito na Tabela 2, computamos a abrangência dos 8 métodos analisados. A Figura 1(a) apresenta o resultado para o evento Airfrance, que relata o trágico acidente de avião em 2009. Como podemos perceber na figura, os métodos SentiWordNet e SenticNet obtiveram a maior abrangência nesse período, com 90% e 91% respectivamente, seguido pelo SentiStrength, com 61%. Emotions e o PANAS-t conseguiram capturar menos de 10% dos tweets relevantes do evento.

No caso das eleições norte-americanas de 2008, a Figura 1(d) mostra que SentiWordNet, SenticNet e SASA tiveram as maiores porcentagens de abrangência, com 90%, 88% e 67%, respectivamente. De fato, SentiWordNet e SenticNet foram ambos os métodos com a maior abrangência verificada em todos os eventos da Tabela 2, intercalando entre eles a primeira posição. Nos outros eventos, SentiStrength, LIWC e SASA ficaram na terceira e quarta posição.

Esses resultados também mostram que, apesar de poucos métodos alcançarem altas abrangências, a porcentagem de tweets não identificados é significativa para a maioria deles. Essa porcentagem pode representar o erro do método na detecção de sentimentos. Uma segunda análise feita verifica a fração de tweets que podem ser identificados se combinarmos mais de 1 método. Para cada evento, combinamos todos os métodos 1 a 1, iniciando pelo que obteve a maior até o que obteve menor abrangência. Ao combinarmos 2 métodos, fomos capazes de aumentar a abrangência em mais de

<sup>2</sup>Eventos Destaques do Twitter em <http://tinyurl.com/yb4965e>

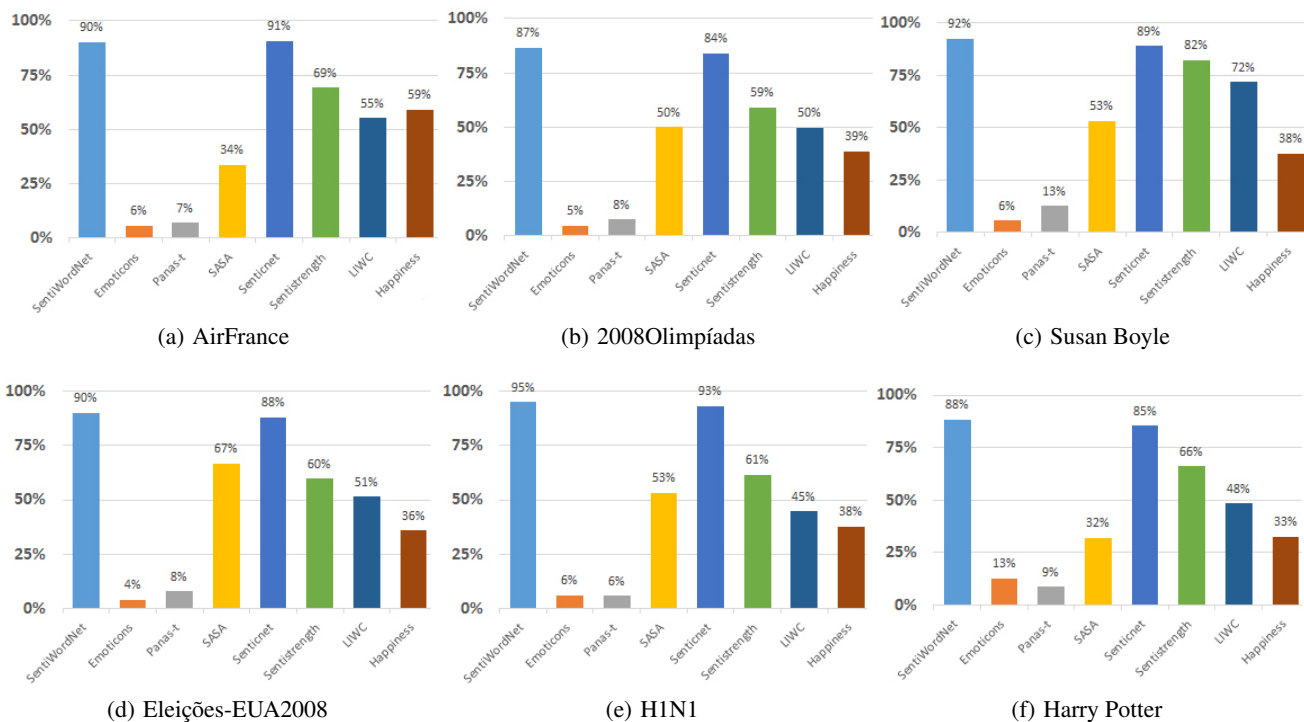


Figura 1: Abrangência dos 6 eventos.

Tabela 4: Porcentagem de concordância entre os métodos

Métrica	PANAS-t	Emoticons	SASA	Sentic-Net	Senti-WordNet	Happiness Index	Senti-Strength	LIWC	Média
PANAS-t	-	60,00	66,67	30,77	56,25	-	74,07	80,00	52,53
Emoticons	33,33	-	64,52	64,00	57,14	58,33	72,00	75,00	60,61
SASA	66,67	64,52	-	64,29	60,00	64,29	61,76	68,75	64,32
SenticNet	30,77	60,00	64,29	-	64,29	59,26	63,33	73,33	59,32
SentiWordNet	56,25	57,14	60,00	64,29	-	64,10	52,94	62,50	59,04
Happiness Index	-	58,33	64,29	62,50	70,27	-	65,52	71,43	56,04
SentiStrength	74,07	75,00	63,89	63,33	52,94	65,52	-	75,00	66,67
LIWC	80,00	75,00	68,97	73,33	58,82	83,33	75,00	-	73,49
Média	48,72	63,85	64,65	60,35	59,95	56,40	66,37	72,29	-

92,75% em todos os eventos. Também notamos que, utilizando essa estratégia, a porcentagem de tweets não identificados foi menor que 7,24% para todos os eventos. Esses resultados são importantes pois conseguimos mostrar que podemos atingir melhores abrangências quando combinamos vários métodos.

## 4.2 Concordância

A seguir, examinamos o grau com que diferentes métodos concordam na polaridade de um conteúdo. Por exemplo, quando 2 ou mais métodos detectam sentimentos em uma mesma mensagem, pode ser importante checar se esses sentimentos foram os mesmos, o que poderia aumentar a confiança da classificação.

A Tabela 4 apresenta a porcentagem da concordância de cada método com todos os outros. Para cada método na primeira coluna, calculamos a fração de mensagens que concordaram com cada método na primeira linha, aos pares. Nossos resultados sugerem que alguns métodos atingem alto grau de concordância quando combinados, como é o caso do LIWC e PANAS-t (80%), enquanto outros possuem pouca concordância, como é o caso do SenticNet e PANAS-t (30,77%). PANAS-t e Happiness Index não concorda-

ram em nenhuma detecção. A última linha da tabela apresenta a média da concordância de cada método com os outros 7. Podemos perceber que o método com maior concordância com outros foi o LIWC, sugerindo que este consiste de um método interessante para ser combinado com outros.

Esses resultados indicam que os métodos variam muito em termos de concordância em que predizem polaridade, variando entre 33% e 80%. Isso implica que, para uma mesma base de dados, a escolha de métodos para detecção de sentimentos pode resultar em diferentes observações. Em particular, para aqueles métodos em que a concordância foi menor que 50%, a polaridade sempre mudará (de positivo para negativo, ou vice versa).

## 4.3 Capacidade da Predição

A seguir, apresentamos uma análise da capacidade de predição de cada método em termos de predição correta de polaridade. Mostraremos resultados para precisão, *recall*, acurácia e F-measure. Para computar essas métricas, utilizamos a base rotulada do SentiStrength [22] para positivo e negativo descrita em §3.2.1.

Para comparar os resultados de performance de predição para cada

**Tabela 5: Média da performance de predição dos métodos para a base rotulada**

Metric	PANAS-t	Emoticons	SASA	Sentic-Net	Senti-WordNet	Happiness Index	Senti-Strength	LIWC
Recall	0,614	0,856	0,648	0,562	0,601	0,571	0,767	0,153
Precision	0,741	0,867	0,667	0,934	0,786	0,945	0,780	0,846
Accuracy	0,677	0,817	0,649	0,590	0,643	0,639	0,815	0,675
F-measure	0,632	0,846	0,627	0,658	0,646	0,665	0,765	0,689

**Tabela 6: F-measures para os 8 métodos**

Método	Twitter	MySpace	Youtube	BBC	Digg	Runners World
PANAS-t	0,643	0,958	0,737	0,296	0,476	0,689
Emoticons	0,929	0,952	0,948	0,359	0,939	0,947
SASA	0,750	0,710	0,754	0,346	0,502	0,744
SenticNet	0,757	0,884	0,810	0,251	0,424	0,826
SentiWordNet	0,721	0,837	0,789	0,284	0,456	0,789
SentiStrength	0,843	0,915	0,894	0,532	0,632	0,778
Happiness Index	0,774	0,925	0,821	0,246	0,393	0,832
LIWC	0,690	0,862	0,731	0,377	0,585	0,895

método, apresentamos a Tabela 5, com a média dos resultados obtidos para cada base rotulada. É importante lembrar que o F-measure alcança seu melhor valor em 1 e pior em 0. Podemos observar que o método com melhor F-measure foi o Emoticons (0,846), porém o de menor abrangência, como já verificado. O segundo melhor método em relação ao F-measure foi o SentiStrength, que obteve uma abrangência muito maior se comparado ao método Emoticons. É importante notar que a versão que estamos utilizando do SentiStrength foi treinada com os dados rotulados utilizados nessa análise a execução de experimentos deste método nessa base de dados poderia potencialmente apresentar viés. Ao invés disso, computamos as métricas de predição para o SentiStrength baseados em valores reportados em experimentos dos próprios autores [22].

A Tabela 6 apresenta o F-measure calculado para cada método para cada base rotulada. Podemos perceber que os 8 métodos apresentam variações nos seus resultados ao longo das diferentes bases de dados, resultando em diferentes performances quando expostos a bases informais (ex.: Twitter e MySpace) ou formais (ex.: BBC e Digg). Por exemplo, para mensagens da BBC, o método SentiStrength foi o com maior F-measure (53%). Por outro lado, para a base de dados do MySpace o maior F-measure foi obtido pelo PANAS-t, e a média dos F-measures de todos os métodos nessa base foi de 72%.

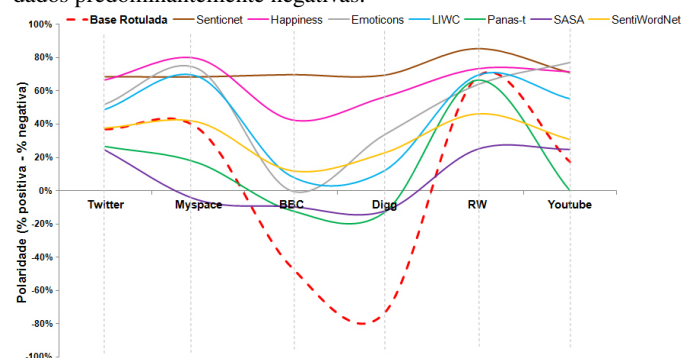
#### 4.4 Análise de Polaridade

Até o momento, analisamos a abrangência e a performance de predição para os métodos. A seguir fornecemos uma análise mais profunda sobre como a polaridade varia em diferentes bases de dados e apresentamos potenciais “armadilhas” que precisamos ter cuidado ao monitorar e medir polaridade em redes sociais.

A Figura 2 apresenta a polaridade dos métodos quando expostos a diferentes bases de dados. Para cada base e método, calculamos a porcentagem de mensagens positivas e negativas identificadas. O eixo-Y mostra a porcentagem de mensagens positivas subtraídas das negativas. Apresentamos no mesmo gráfico uma curva que representa o gabarito para essa análise, possível em razão da base de dados rotulada utilizada nessa análise. Assim, quando mais próximo a essa curva, melhor foi a predição de polaridade do método. SentiStrength foi desconsiderado dessa análise já que o mesmo foi treinado utilizando a base de dados envolvida no processo de construção do gráfico.

A partir dos resultados obtidos, podemos realizar diversas observa-

ções interessantes. Primeiramente, podemos perceber que a maioria dos métodos tendem a apresentar mais sentimentos positivos que negativos, já que observamos poucas curvas abaixo daquela que representa o gabarito para a base. Segundo, podemos notar que muitos métodos obtiveram apenas valores positivos, independente da base de dados analisada. Uma observação interessante é o fato de o método SenticNet ter apresentado as maiores taxas de abrangência, porém identificou polaridades incorretas para conjuntos de dados predominantemente negativos.

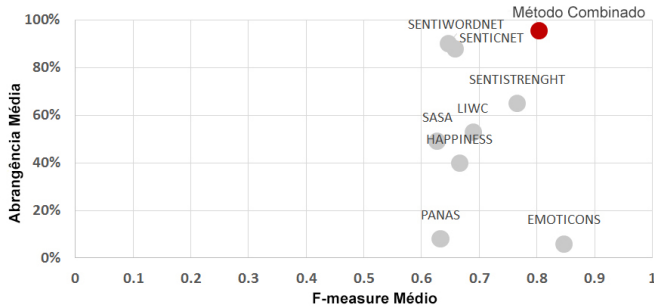


**Figura 2: Polaridade dos 8 métodos**

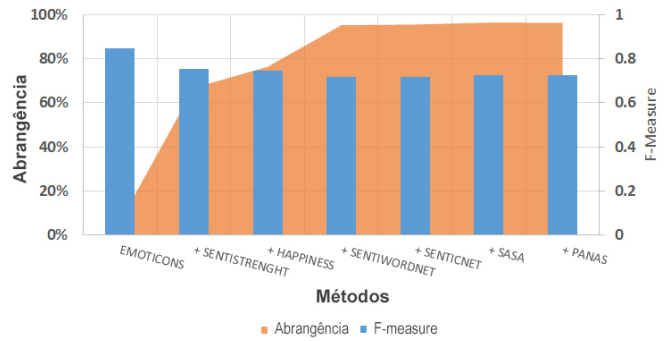
O viés na identificação de sentimentos positivos apresentado pela maioria dos métodos pode atrapalhar a detecção em tempo real de ferramentas desenvolvidas para esse contexto, já que essas simplesmente aplicam esses métodos em dados coletados e calculam a taxa de mensagens positivas e negativas nos textos. Os resultados poderiam potencialmente conter viés devido ao método utilizado. Com objetivo de verificar esse viés, realizamos o mesmo tipo de análise para cada evento filtrado do Twitter. A partir dessa análise, verificamos que a maioria dos métodos apresentam resultados positivos mesmo em eventos como H1N1, do qual esperávamos uma maior quantidade de tweets expressando sentimentos negativos. Da mesma forma, o evento Airfrance foi considerado positivo por 4 métodos, embora a polaridade tenha ficado próxima de zero para a maioria deles.

#### 5. MÉTODO COMBINADO

Tendo em vista os diferentes graus de abrangência e acurácia dos 8 métodos para detecção de sentimentos analisados, apresentamos



(a) Comparação



(b) Combinação incremental

Figura 3: Abrangência vs. F-measure, para todos os métodos

um método que consiste da combinação destes métodos.

## 5.1 Método Combinado

Construímos um método que consiste na combinação de 7 dos 8 métodos analisados, ao qual damos o nome de **Método Combinado**. Esse novo método inclui os seguintes métodos: PANAS-t, Emoticons, SentiStrength, SentiWordNet, SenticNet, SASA e Happiness Index. O LIWC foi omitido do processo de construção do método por razões de restrições de cópia. O Método Combinado analisa a média harmônica (F-measure) da precisão e *recall* de cada método e distribui diferentes pesos para cada um deles.

Para avaliação, testamos nosso método sobre a base de dados do SentiStrength [22] que consiste de mensagens rotuladas por humanos do AMT (veja descrição na §3.1.2). Calculamos o F-measure e a média de abrangência em cima desses dados. Também computamos a abrangência baseada na base de dados que consiste em um histórico completo do Twitter, utilizando a média de abrangência através dos 6 eventos analisados (veja descrição na §3.1.1).

A Figura 3(a) compara a abrangência e F-measure dos 7 métodos utilizados nesta seção, assim como do Método Combinado. O resultado mostra a eficácia do Método Combinado, que conseguiu alcançar abrangência de 95%, como esperado. A acurácia e precisão do método também se manteve relativamente alta, com F-measure igual a 0,730, valor menor que o método com melhor performance, o Emoticons, porém maior que todos os outros métodos analisados.

Enquanto combinar todos os métodos de detecção de sentimentos pode melhorar a abrangência, há apenas um pequeno ganho marginal dessa quando começamos a aumentar o número de métodos na análise. A Figura 3(b) apresenta essa característica, onde adicionamos métodos na ordem Emoticons, SentiStrength, Happiness Index, etc (apresentada no eixo-X). Enquanto o método Emoticons nos dá a menor abrangência (menos de 10%), esta começa a aumentar (70%) quando adicionamos o método seguinte, nesse caso o SentiStrength (região alaranjada da figura). O F-measure, por outro lado, decresce levemente como podemos ver nas barras em azul na figura, e continua a cair na medida com que métodos novos são adicionados.

A medida que combinamos mais métodos, a abrangência aumenta pouco. De fato, combinando os primeiros 4 métodos atingimos uma abrangência de 95%, deixando pouco espaço para melhoras após esse ponto. Podemos também notar que, apesar da acurácia e precisão decrescer, assim que mais métodos são combinados, o F-measure ainda permanece acima dos 0,7. Isso pode indicar que combinar todos os métodos não necessariamente traz os melhores resultados. Ao invés disso, melhores resultados poderiam ser al-

cançados escolhendo o conjunto de métodos que melhor se ajusta ao que o usuário necessita. Dessa forma, poderíamos desenvolver aplicações reais envolvendo poucos métodos, porém com bons resultados.

## 5.2 iFeel

Este trabalho permitiu o desenvolvimento de um sistema Web que chamamos de iFeel. O iFeel consiste de uma ferramenta que permite a comparação do resultado de detecção de sentimentos para vários métodos facilmente. O sistema também dá acesso aos resultados do Método Combinado, que tipicamente dará melhor abrangência e acurácia. Acreditamos que esse sistema é muito útil, já que poderá auxiliar pesquisadores e empresas na análise de sentimentos em dados particulares. A Figura 4 apresenta uma *screenshot* do funcionamento do iFeel para a mensagem de teste "I'm feeling too sad today :(".

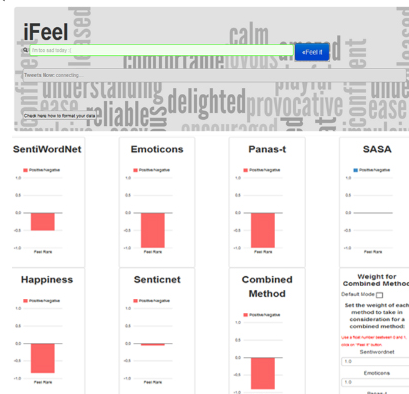


Figura 4: Screenshot da ferramenta iFeel

O iFeel está disponível para acesso em [www.ifeel.dcc.ufmg.br](http://www.ifeel.dcc.ufmg.br).

## 6. CONCLUSÕES

Pesquisas recentes em redes sociais online vêm adotando vários métodos para análise de sentimentos em conteúdo postado na Web. Vários desses se tornaram populares e estão sendo utilizados em ferramentas para medir a polaridade em redes sociais online. Nesse artigo, apresentamos diversas comparações entre 8 métodos muito utilizados: SentiWordNet, SASA, PANAS-t, Emoticons, SentiStrength, LIWC, SenticNet, e Happiness Index.

Nossos estudos de comparações focaram na detecção da polaridade (positivo ou negativo) em conteúdo da Web, porém ainda não con-

siderou outros tipos de sentimentos (ex.: afetos psicológicos como raiva, calma, etc). Adotamos várias métricas para medir a eficácia de um método: abrangência (medindo a fração de mensagens que foram capturados por um método), concordância (medindo a concordância entre a polaridade entre os métodos utilizando uma base de dados rotulada), taxa de *true positive* (taxa em que mensagens positivas foram corretamente identificadas), taxa de *true negative* (taxa em que mensagens negativas foram preditas como negativas), acurácia (taxa em que um método identificou sentimentos corretamente), precisão (calcula o quão próximo os valores medidos estão um do outro) e F-measure, que relaciona precisão e *recall*. Como resultados, percebemos que os 8 métodos possuem variados graus de abrangência e acurácia, e não existe um método com melhores resultados sempre. Esses resultados nos levou a construção de um novo método que consiste da combinação dos outros na tentativa de alcançar melhores abrangências e acurácia satisfatória. Esse método foi apresentado com nome Método Combinado.

Neste trabalho, apresentamos um cenário de comparação entre vários métodos para análise de sentimentos. Para realizar esta tarefa, cobrimos um amplo conjunto de pesquisas em análise de sentimentos e realizamos um esforço significativo em contatar seus respectivos autores para ter acesso aos métodos. Infelizmente, em muitos casos, ter acesso aos métodos não foi fácil e portanto realizamos estudos com base em apenas 8 métodos. Como trabalhos futuros pretendemos incorporar mais métodos existentes na literatura para comparação, como o *Profile of Mood States* (POMS) [7] e *OpinionFinder* [26]. Além disso, gostaríamos de expandir as categorias de sentimentos comparados para além de polaridade positivo e negativo.

## 7. AGRADECIMENTOS

Esse trabalho teve apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG), da da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e também do Instituto Nacional de Ciência e Tecnologia para a Web (InWeb).

## 8. REFERENCES

- [1] Msn messenger emoticons. <http://messenger.msn.com/Resource/Emoticons.aspx>.
- [2] Omg! oxford english dictionary grows a heart: Graphic symbol for love (and that exclamation) are added as words. [tinyurl.com/klv36p](http://tinyurl.com/klv36p).
- [3] Sentistrength 2.0. <http://sentistrength.wlv.ac.uk/Download>.
- [4] Yahoo messenger emoticons. <http://messenger.yahoo.com/features/emoticons>.
- [5] Amazon. Amazon mechanical turk. <https://www.mturk.com/>. Accessed June 17, 2013.
- [6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.
- [7] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.
- [8] M. M. Bradley and P. J. Lang. Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.
- [9] E. Cambria, A. Hussain, C. Havasi, C. Eckl, and J. Munro. Towards crowd validation of the uk national health service. In *ACM Web Science Conference (WebSci)*, 2010.
- [10] E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium Series*, 2010.
- [11] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.
- [12] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2009.
- [13] Esuli and Sebastiani. Sentwordnet: A publicly available lexical resource for opinion mining. In *In Conference on Language Resources and Evaluation*, 2006.
- [14] P. Goncalves and F. Benevenuto. O que tweets contendo emoticons podem revelar sobre sentimentos coletivos? In *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2013.
- [15] P. Goncalves, W. Dores, and F. Benevenuto. Panas-t: Uma escala psicometrica para analise de sentimentos no twitter. In *I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2012.
- [16] A. Hannak, E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.
- [17] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [18] J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.
- [19] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACL Student Research Workshop*, pages 43–48, 2005.
- [20] S. Somasundaran, J. Wiebe, and J. Ruppenhofer. Discourse level opinion interpretation. In *Int'l Conference on Computational Linguistics (COLING)*, pages 801–808, 2008.
- [21] 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):24–54, 2010.
- [22] M. Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. <http://migre.me/fHgJ9>.
- [23] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *ACL System Demonstrations*, 2012.
- [24] D. Watson and L. Clark. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(1):1063–1070, 1985.
- [25] K. Wickre. Celebrating twitter7. <http://migre.me/fHgJ9>.
- [26] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *HLT/EMNLP on Interactive Demonstrations*, 2005.