Detecção semi-supervisionada de posicionamento em tweets baseada em regras de sentimento

Marcelo Dias, Karin Becker

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

{marcelo.dias,karin.becker}@inf.ufrgs.br

Resumo. *A detecção de posicionamento visa identificar automaticamente se o autor de um texto é a favor ou contra um dado alvo. O presente trabalho descreve um método semi-supervisionado de detecção de posicionamento no conteúdo textual de tweets. Um conjunto de regras é proposto para identificar posicionamento com base em opiniões positivas ou negativas a alvos direta ou indiretamente relacionados. Os tweets rotulados pelas regras são utilizados para compor um corpus de treinamento para uma abordagem supervisionada. O modelo preditivo resultante complementa a rotulação feita usando as regras. O artigo apresenta o método, e uma análise de seu desempenho quando aplicado a diferentes domínios, como candidatos políticos, mudança do clima e aborto.*

Abstract. *Stance detection aims to automatically identify if the text author is in favor or against a subject or target. This work describes a semi-supervised method for stance detection. The core is a set of rules to identify stance based on positive or negative opinions of targets directly or indirectly related. Tweets automatically labeled using the rules compose a training corpus for a supervised approach. The resulting predictive model allows to predict the stance of unlabeled tweets. This paper presents the method and analyzes the obtained results when applied to different data domains like political candidatures, climate change and legalization of abortion.*

1. Introdução

Twitter tornou-se chave na disseminação de ideias, opiniões, crenças, e posicionamentos sobre os mais diversos assuntos. A identificação automática de estados afetivos diversos é o objetivo da análise de sentimentos [Liu, 2012]. Detecção de posicionamento é a tarefa de identificar automaticamente se o autor de um texto é a favor, contra ou neutro em relação a um alvo. O alvo do posicionamento pode ser uma entidade concreta, como uma pessoa, organização ou local, ou abstrata, como uma causa ou afirmação. Esta área revela semelhanças e diferenças em relação à análise de opiniões, que busca medir a valência ou polaridade do sentimento em relação ao alvo da opinião. Contudo, um dos principais desafios da detecção de posicionamento é a relação entre o alvo da opinião e o alvo do posicionamento, e de como este sentimento revela apoio ou oposição. Por exemplo, se o alvo do posicionamento for um candidato político, sentimentos positivos ou negativos em relação a seu partido, seus apoiadores ou opositores, ou elementos de seu programa de governo são uma forma de revelar o posicionamento.

Trabalhos na área de detecção de posicionamento são baseados em documentos bem estruturados [Thomas et al., 2006, Faulkner, 2014], ou debates [Somasundaran e Wiebe, 2009, Anand et al., 2011]. Estes trabalhos utilizam as características textuais do texto, onde o contexto é mais esclarecedor, ou relações estruturais entre

elementos de um debate. A detecção de posicionamento em tweets é uma área relativamente nova. As características da plataforma social (e.g. perfis, retweets) são utilizadas em [Rajadesingan e Liu, 2014] como forma de dar contexto ao texto dos tweets. O 2016 International Workshop on Semantic Evaluation (SEMEVAL)[1] lançou uma competição de detecção de posicionamento em tweets baseada exclusivamente no conteúdo textual dos tweets [Mohammad et al., 2016].

O presente trabalho descreve uma abordagem semi-supervisionada voltada à detecção de posicionamento em tweets, juntamente com a análise dos resultados quando aplicada a diferentes domínios. A detecção de posicionamento em tweets é feita baseada exclusivamente no conteúdo dos textos, sem fazer uso de informações estruturadas subjacentes às plataformas. A premissa é de que o problema de detecção de posicionamento é um problema com um forte componente de detecção de polaridade, onde o principal desafio é relacionar o alvo da opinião e do sentimento em relação a este com o alvo principal do problema. A abordagem consiste em rotular automaticamente um conjunto de instâncias a fim de compor um corpus de treinamento para um método supervisionado de aprendizado. A rotulação automática é feita com base em regras que combinam alvos, sentimento e expressões que caracterizem apoio ou oposição. A abordagem é semi-supervisionada pois requer como entrada um conjunto de n-gramas que representem, no domínio, alvos relacionados ou expressões (e.g. hashtags) que caracterizem apoio ou oposição. Uma versão preliminar deste trabalho [Dias e Becker, 2016] foi submetida à competição de detecção não-supervisionada de posicionamento do SEMEVAL, obtendo a terceira colocação, mas se limitou à detecção de posicionamento sobre um único alvo (i.e. Trump).

Neste artigo, analisamos a adequação desta abordagem para a detecção de posicionamento considerando domínios distintos: política, feminismo, legalização do aborto, ateísmo e mudança do clima. Além da discussão dos resultados obtidos para cada conjunto de dados referente a estes domínios, analisamos a contribuição e qualidade das regras propostas relacionadas aos resultados, já que têm o objetivo de representar nossas premissas. A abordagem alcançou resultados promissores, particularmente para posicionamento contrário (medida F de 60%), e mostrou a generalidade das regras propostas para domínios com diferentes características.

Em relação aos trabalhos relacionados, as principais contribuições deste trabalho são: a) apresentar uma abordagem de detecção de posicionamento tendo como base exclusivamente o conteúdo de tweets [Thomas et al., 2006, Faulkner, 2014, Somasundaran e Wiebe, 2009, Anand et al., 2011]; b) oferecer uma alternativa às tradicionais abordagens supervisionadas que requerem grande quantidade de dados rotulados [Rajadesingan e Liu, 2014]; c) demonstrar a generalidade da solução quando aplicada a diversos domínios [Dias e Becker, 2016].

O restante do artigo está organizado como segue. A Seção 2 discute trabalhos relacionados. A Seção 3 detalha a abordagem proposta para detecção semi-supervisionada de posicionamento em tweets. A Seção 4 discute os experimentos envolvendo seis diferentes alvos. Conclusões e trabalhos futuros são apresentados na Seção 5.

2. Trabalhos relacionados

A mineração de opiniões em redes sociais como o Twitter, é uma área bastante explorada. O aprendizado supervisado costuma apresentar resultados superiores às abordagens

[1] http://alt.qcri.org/semeval2016/task6/

léxicas [Liu, 2012], mas requer um extenso corpus anotado no domínio. A rotulação automática, por exemplo, baseada em emoticons, é uma solução bastante explorada como forma de contornar este problema (e.g. [Tang et al., 2015]).

A detecção de posicionamento é um tema menos explorado. Os trabalhos pioneiros nesta área consideram principalmente debates, tais como a análise de debates políticos no congresso dos Estados Unidos da América [Thomas et al., 2006] e ocorridos em ambientes para debates [Somasundaran e Wiebe, 2009][Anand et al., 2011]. Há também propostas que analisam trabalhos acadêmicos [Faulkner, 2014]. Nestes casos, a detecção de posicionamento ocorre em um texto mais completo, e pode se valer de informações estruturadas, tais como o voto do autor do texto, a refutação de um comentário anterior no debate, ou o posicionamento informado pelo próprio autor junto ao texto.

A detecção de posicionamento em tweets propõe novos desafios, pois os textos são menores, as threads difíceis de serem identificadas, além dos problemas usuais de vocabulário e mau uso da linguagem. Um método semi-supervisionado de detecção de posicionamento em tweets é descrito em [Rajadesingan e Liu, 2014], que utiliza informações sociais disponíveis na plataforma Twitter. A abordagem é dependente da capacidade de reconhecimento de perfis notórios (i.e. reconhecidamente a favor ou contra), os quais são utilizados para propagação de rótulos a tweets de seguidores, visando formar um corpus para a criação de um classificador. O presente trabalho diferencia-se deste por não depender da identificação de perfis notórios, que pode ser complexa para alguns domínios, e por incluir o posicionamento neutro no processo de classificação.

O SEMEVAL 2016 propôs duas tarefas para identificação de posicionamento baseada exclusivamente no texto dos tweets, supervisionada e não supervisionada, mas detalhes das abordagens submetidas não são ainda conhecidos [Mohammad et al., 2016]. Uma versão preliminar do presente trabalho foi submetida à tarefa de identificação não supervisionada, obtendo terceiro lugar dentre nove participantes [Dias e Becker, 2016]. Contudo ela se limitou a um único alvo, a saber, o candidato político Trump. Estendemos este trabalho através de ajustes à abordagem, bem como de uma análise de seu desempenho face a seis diferentes domínios.

3. Visão Geral da Abordagem

A Figura 1 destaca os principais pontos da abordagem proposta para detecção semi-supervisionada de posicionamento baseado no conteúdo de tweets. Ela compreende dois aspectos: a) criação de um modelo preditivo a partir da criação automática de um corpus de treinamento, e b) detecção de posicionamento. A abordagem é semi-supervisionada porque pressupõe como entrada um pequeno conjunto de n-gramas que representem os alvos de opinião direta ou indiretamente relacionados, ou expressões explícitas de posicionamento (e.g. hashtags ou tópicos tipicamente mencionados por um dos lados no domínio). No restante da seção, são discutidas as principais atividades necessárias para identificação de posicionamento em conteúdos de tweets.

3.1. Criação de um Modelo Preditivo

3.1.1. Identificação de N-gramas Alvo e Chave

A criação de um modelo preditivo inicia com a identificação de um conjunto de n-gramas (uni-gramas, bi-gramas e tri-gramas) que sejam representativos para expressar posicionamentos no domínio, os quais dividimos em dois grandes grupos: *alvo* e *chave*.

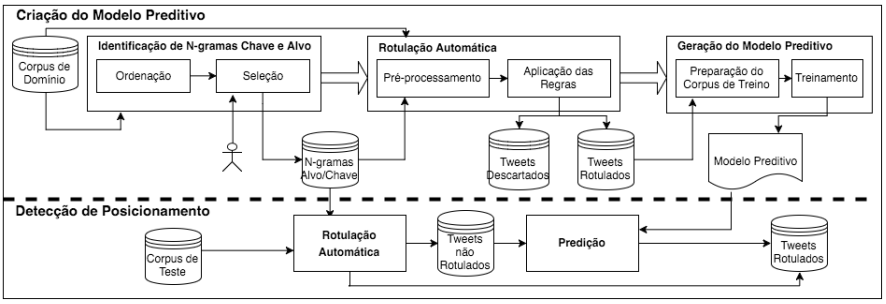


Figura 1. Pocesso de Detecção de Posicionamento em Tweets

Os n-gramas alvo são aqueles utilizados em tweets para representar direta ou indiretamente alvos a favor da questão, ou o lado opositor. Por exemplo, para o problema de posicionamento em relação a politicos (e.g. Hillary Clinton), alvos do lado favorável são variações do nome do politico (e.g. Hillary, Clinton), seus aliados (e.g. Bill, Obama), ou seu partido (e.g. democratas), enquanto para o lado contrário poderíamos citar seus concorrentes (e.g. republicanos, Trump, Bush). Além de entidades, há também assuntos caracterizando alvos, tais como elementos do programa de governo. Por exemplo, opiniões sobre "imigrantes" ou "mexicanos" caracterizam posicionamento sobre Trump, enquanto que "saúde pública" é utilizado para se posicionar em relação à Hillary.

Já os n-gramas chave são normalmente hashtags ou expressões que, quando mencionadas, denotam uma forte tendência quanto ao posicionamento do tweet. Exemplos de n-gramas de posicionamento favorável/contrário à Hillary são "readyforhillary" e "stophillary2016", respectivamente.

N-gramas alvo e chave devem ser separados em 2 subgrupos distintos: *Favoráveis* e *Contrários*, considerando o alvo principal. Por exemplo, "hillary" é um n-grama alvo favorável ao alvo principal Hillary, enquanto que "trump" é um n-grama alvo contrário.

Propõe-se que os n-gramas sejam extraídos de um conjunto de tweets no domínio (identificado na Figura 1 como *Corpus de Dominio*), em um processo semi-automático. Um procedimento extrai deste corpus os n-gramas mais frequentes, para um limiar dado, e os ordena por frequência. Esta lista é apresentada a um anotador para inspeção manual, de forma que possa analisá-los, selecionando os mais relevantes.

3.1.2. Rotulação Automática de Tweets

A rotulação automática de tweets permite compor um corpus de treinamento para aprendizagem supervisionada, evitando o custo associado a esta atividade quando executada de forma manual. Para este fim, propusemos um conjunto de regras, apresentadas na Tabela 1, as quais representam nossa premissa básica de que posicionamentos podem ser expressos na forma de sentimentos positivos ou negativos em relação a alvos direta ou indiretamente relacionados. As regras 1 e 2 inferem o posicionamento do tweet com base na presença de n-gramas chave, enquanto que as regras de 3 a 6 rotulam o tweet com base na presença de n-gramas de alvo combinado com a polaridade detectada para o texto do Tweet. A regra 7 assume que não há posicionamento para tweets onde não foi detectado sentimento. Dado o tamanho de um tweet, assume-se que o alvo de um sentimento é um n-grama alvo.

Esta etapa é dividida em dois passos: *pré-processamento*, que gera as informações

Regra	Descrição	Posicionamento
1 - CHAVE-FAVOR	Presença de n-grama(s) chave favorável(eis) e ausência de n-grama chave contrário	FAVORÁVEL
2 - CHAVE-CONTRA	Presença de n-grama(s) chave contrário(s) e ausência de n-grama chave favorável	CONTRÁRIO
3 - ALVO-FAVOR-POSITIVO	Presença de n-grama(s) alvo favorável(eis), ausência de n-grama contrário e polaridade positiva	FAVORÁVEL
4 - ALVO-FAVOR-NEGATIVO	Presença de n-grama(s) alvo favorável(eis), ausência de n-grama contrário e polaridade negativa	CONTRÁRIO
5 - ALVO-CONTRA-POSITIVO	Presença de n-grama(s) alvo contrário(s), ausência de n-grama favorável e polaridade positiva	CONTRÁRIO
6 - ALVO-CONTRA-NEGATIVO	Presença de n-grama(s) alvo contrário(s), ausência de n-grama favorável e polaridade negativa	FAVORÁVEL
7 - NEUTRO	Tweet de polaridade neutra	SEM POSIC.
Outros casos		DESCARTAR

Tabela 1. Regras para Rotulação Automática

necessárias para aplicação das regras, e a *aplicação das regras* propriamente dita. Como resultado, tem-se dois conjuntos de tweets: um rotulado com o posicionamento assumido (FAVORÁVEL, CONTRÁRIO ou NENHUM), e um de tweets descartados, i.e. que não foram filtrados por nenhuma regra. Dado o conjunto de n-gramas chave e alvo, o passo de pré-processamento tem por objetivo gerar as seguintes informações para cada tweet do Corpus de Domínio: a) Presença de pelo menos um n-grama chave favorável ao alvo principal; b) Presença de pelo menos um n-grama chave contrário ao alvo principal; c) Presença de pelo menos um n-grama alvo favorável ao alvo principal; d) Presença de pelo menos um n-grama alvo contrário ao alvo principal e e) Polaridade do Tweet.

3.1.3. Criação de um Modelo Preditivo de Posicionamento Utilizando Aprendizado Supervisionado

Considerando o texto dos tweets do Corpus do Domínio, e os respectivos rótulos atribuídos automaticamente na etapa anterior, o objetivo desta etapa é compor um corpus de treinamento, e usá-lo para treinar um classificador visando a geração de um modelo preditivo. Como já salientado, apenas os tweets que foram filtrados pelas regras 1 a 7 da Tabela 1 são considerados para o corpus de treinamento. Em nossos experimentos, esta etapa foi realizada utilizando um algoritmo de classificação (SVM).

Em termos de processamento textual dos tweets, a criação do corpus de treinamento implica as seguintes ações: a) substituição de menções a perfis de twitter pelo unigrama `"a_mention"`, desde que não estejam relacionados aos n-gramas alvo identificados na primeira etapa do processo (e.g. "realdonaldtrump"), b) remoção de stopwords, c) extração de todos uni-gramas, e d) criação do corpus usando peso binário para os atributos (i.e. presença ou ausência do uni-grama no texto).

3.2. Identificação de Posicionamento em Tweets não Rotulados

Nossa abordagem identifica o posicionamento contido no texto de tweets combinando a rotulação automática, no exato processo descrito na Seção 3.1.2, com o modelo preditivo gerado através do processo descrito na Seção 3.1.3. Dado um corpus contendo um conjunto de tweets não rotulados, representado na Figura 1 como *Corpus de Teste*, primeiro tenta-se rotulá-los com o apoio das regras da Tabela 1. Somente os tweets que não são filtrados por nenhuma das regras (*Tweets Não Rotulados*) são submetidos ao modelo preditivo. Este processo difere da proposta original [Dias e Becker, 2016], pois testes demonstraram que as regras têm melhor precisão que o modelo preditivo, mas filtram apenas uma pequena parcela dos tweets, como discutido na Seção 4.

4. Experimento e Resultados

Com a intenção de desenvolver uma abordagem abrangente para a detecção de posicionamento em tweets, utilizamos todos os conjuntos de dados disponibilizados pelo SE-

MEVAL 2016 para as tarefas de detecção supervisionada e não supervisionada de posicionamento em tweets[2]. Nosso método foi originalmente desenvolvido considerando as características de um único alvo de posicionamento (o candidato Donald Trump), e através destes experimentos avaliamos sua adequação a diferentes domínios.

4.1. Ambiente Experimental

Para realização dos experimentos desenvolvemos programas específicos para auxiliar na seleção de n-gramas e para realizar a rotulação automática. Para a criação do modelo preditivo, foi utilizada a plataforma Weka (versão 3.7.11). A implementação SMO do algoritmo SVM foi escolhida após realizarmos testes com outros algoritmos, pois apresentou resultado superior a outras opções testadas (i.e. Random Forest, RBF, Naive Bayes, e meta-classificador Vote para formação de um comitê de classificadores).

Para a detecção de polaridade, utilizamos uma combinação de APIs de análise de sentimento, a saber HP Haven On Demand[3], IBM Alchemy[4] e Vivekn[5]. Cada API retorna a polaridade detectada e uma propriedade que representa um escore. O programa verifica inicialmente se a polaridade é neutra para as APIs Haven e Alchemy e, se ambas tiverem valor neutro, considera o tweet como neutro. Caso contrário, a propriedade escore retornada pelas APIs é somada, classificando o tweet como positivo se a soma resultar um número maior que zero, e negativo caso a soma resulte em um número menor que zero. Utilizando um conjunto de dados do domínio político rotulado para polaridade [6] [Mohammad et al., 2015], comparamos o desempenho das 3 APIs em isolado com o da combinação proposta [Dias e Becker, 2016]. Obtivemos como medida F ponderada 71% para a polaridade negativa, 61.5% para a polaridade positiva, e apenas 20.4% para a polaridade neutra. O maior benefício obtido foi uma significativa melhora no desempenho para a classe positiva em mais de 25 pontos percentuais, já que os resultados para as classes negativa e neutra são muito similares ao uso da API Alchemy de forma isolada.

4.2. Caracterização dos Corpora de Domínio

Utilizamos 6 conjuntos de dados disponibilizados no SEMEVAL 2016: a) Ateísmo, b) Mudança do Clima é uma preocupação real, c) Movimento Feminista, d) Legalização do Aborto, e) Hillary Clinton e f) Donald Trump. Para cada alvo, foram fornecidos um conjunto de treino e um conjunto de teste, este último usado para avaliação da tarefa. Os primeiros 5 conjuntos de dados foram alvo da tarefa de identificação supervisionada, e eram portanto rotulados. Já o conjunto Trump foi disponibilizado para a tarefa de identificação não supervisionada, e portanto apenas o conjunto de teste possui rótulos.

A Tabela 2 mostra a distribuição dos dados em relação ao posicionamento e sentimento, tal como sumarizada pelos organizadores da tarefa[7]. Considerando o conjunto de dados como um todo, o posicionamento *Contrário* ao alvo principal é predominante (49,47%), e o restante distribuído uniformemente nas demais classes (24,74% como Favoráveis, e 25,79% *Sem Posicionamento*). Contudo, ao considerar a distribuição por alvo, existe uma maior ocorrência de instâncias da classe *Sem Posicionamento* do que da classe *Favorável*. Há duas exceções: a) no alvo *Feminismo*, onde a classe *Favorável* supera

[2]http://alt.qcri.org/semeval2016/task6/
[3]https://www.havenondemand.com/
[4]http://www.alchemyapi.com/
[5]http://sentiment.vivekn.com/docs/api/
[6]http://www.purl.org/net/PoliticalTweets2012
[7]http://www.saifmohammad.com/WebPages/StanceDataset.htm

	Instâncias	Posicionamento						Polaridade			
	Total	Contrário		Favorável		Sem Posic.		Negativo		Positivo	
Corpus	**Nro.**	**Nro.**	**%**	**Nro.**	**%**	**Nro.**	**%**	**Nro.**	**%**	**Nro.**	**%**
Aborto	**933**	544	58,31	167	17,9	222	23,70	634	67,95	245	26,26
Ateísmo	**733**	464	63,3	124	16,92	145	19,78	258	35,20	440	60,03
Clima	**564**	26	4,61	335	59,4	203	35,99	283	50,18	175	31,03
Feminismo	**949**	511	53,85	268	28,24	170	17,91	730	76,92	174	18,34
Hillary	**984**	565	57,42	163	16,57	256	26,06	648	65,85	297	30,18
Trump	**707**	299	42,29	148	20,93	260	36,78	481	68,03	193	27,3
Total	**4.870**	2.409	49,47	1.205	24,74	1.256	25,79	3.034	62,30	1.524	31,29

Tabela 2. Instâncias por Posicionamento e Sentimento

Regra	Aborto		Ateísmo		Clima		Feminismo		Hillary		Trump		Média (%)
	Nro.	%	Nro.	%	Nro.	%	Nro.	%	Nro.	%	Nro.	%	
1-CHAVE-FAVOR	21	2	36	5	17	3	103	11	48	5	1	0	4
2-CHAVE-CONTRA	248	27	107	15	0	0	105	11	163	17	1411	7	13
3-ALVO-FAVOR-POSITIVO	7	1	0	0	53	9	28	3	72	8	2350	11	5
4-ALVO-FAVOR-NEGATIVO	97	10	0	0	42	7	195	21	163	17	7497	36	15
5-ALVO-CONTRA-POSITIVO	17	2	97	13	1	0	2	0	8	1	3	0	3
6-ALVO-CONTRA-NEGATIVO	55	6	156	21	18	3	3	0	35	4	43	0	6
7-NEUTRO	56	6	61	8	84	15	58	6	59	6	2272	11	9
OUTROS CASOS	432	46	276	38	349	62	455	48	386	41	7082	34	45
TOTAL	933		733		564		949		934		20659		

Tabela 3. Distribuição das Regras por Corpus de Domínio

a classe *Sem Posicionamento*, e b) no alvo *Clima*, a classe *Favorável* é predominante (59,40%), e a classe *Contrária* é pouco representativa (4,61%)

Em termos de distribuição da polaridade, a classe negativa predomina em todos os alvos, com exceção do alvo *Ateísmo*, onde predominam os tweets positivos (60,03%). No conjunto, os tweets se distribuem em 62,30% na classe negativa, 31,29% na classe positiva, e somente 6,41% na classe neutra. Isso demonstra uma tendência à crítica caracterizada por comentários de polaridade negativa, tanto para expressar oposição, quanto para expressar um posicionamento favorável usando um alvo contrário.

4.3. Seleção de N-gramas Alvo e Chave

A identificação de n-gramas foi realizada considerando a inspeção manual dos 200 n-gramas mais frequentes em cada conjunto de treino. A seleção foi realizada pelos autores utilizando critérios subjetivos, e os n-gramas escolhidos são listados na Tabela 4. Podemos verificar uma menor ocorrência de n-gramas chave favoráveis, além de algumas diferenças entre os alvos quanto à distribuição dos tipos de n-gramas. Alvos que envolvem religião (*Aborto* e *Ateísmo*) e candidatos políticos (*Hillary* e *Trump*) apresentam uma variedade maior de n-gramas chave contrários, o que evidencia o uso de hashtags negativas contra o alvo principal. Já para os alvos *Feminismo* e *Aborto*, percebeu-se uma maior variedade de n-gramas chave favoráveis como forma de expressar apoio à causa (e.g. hashtags "prochoice", "weneedfeminism"). O alvo *Ateísmo* tem um comportamento diferente dos demais, pois não identificamos n-gramas alvo favoráveis, mas sim diversos n-gramas alvo contrários. Com efeito, observamos que opiniões positivas em relação a alvos religiosos são uma forma frequente de manifestar oposição ao ateísmo. Também não identificamos alvos chave contrários para o alvo *Clima*, o que se explica pela tendência de posicionamento favorável neste tema no corpus de treino.

4.4. Análise das Regras

A Tabela 3 mostra a distribuição da aplicação das regras para cada conjunto de dados. Pode-se verificar que as regras cobrem em média 55% das instâncias submetidas ao processo de rotulação automática.

A regra 4 (ALVO-FAVOR-NEGATIVO) é a mais aplicada na maioria dos alvos. Este fato deve-se a duas razões principais. Primeiro, ele ocorre principalmente em alvos

Alvo	N-Gramas Chave		N-Gramas Alvo	
	Favorável	Contrário	Favorável	Contrário
Aborto	womensrights, prochoice, rapeculture, womens rights	unborn children, prolife, love, murder, pro-life, unborn babies, kill, killing, killed, christian, innocent, unborn, prolifegen, human life, alllivesmatter, right to choose, lovewins, prolifeyouth	abort, abortions, abortion	catholic, fetus, kids, the unborn, god, child, baby, children, babies
Ateísmo	freedom freethinker	amen, bless, blessed, god pray, hope, lovewins, mercy, name of jesus, pray for us, sinners, teamjesus, truth	-	bible, book, christ, lord, church, rosary, christian, christianity, mary, holy, spirit, christians, faith, god, gods, heaven, jesus, islam, spirituality, the lord, prayer, religion, religions, religious,
Clima	carbon	-	david attenborough, green, obama, global warming, attenborough, climate, climate change, change, cop21, warming	gop, emissions, co2
Feminismo	weneedfeminism, heforshe, patriarchy, gender, sexist, equal rights, yesallwomen	god, slut, feminazi, spankafeminist, gamergate	females, female, equality, feminism, feminist, girls, feminists, woman, women	family, dear feminists
Hillary	readyforhillary, lovewins, love	bernie2016, liberals, benghazi, whyimnotvotingforhillary, tcot, nohillary2016, stophillary2016, emails, wakeupamerica	bill, clintons, barackobama, uniteblue, democrats, hillarys, billclinton, obama, hillaryclinton hillary clinton, clinton, hillary	bush, republican, gop republicans, marcorubio, trump, realdonaldtrump, berniesanders, bernie,
Trump	stop hillary, stophillary	love wins, lovewins, apprentice, dontvotefortrump, mr trump, racist	realdonaldtrump, donaldtrump, donald trump, donald, trump, republican, republicans	hillaryclinton, hillary, hill, latinos, hillary clinton, hilary, mexican, chapo, hilary clinton,clinton, clintons, obama, democrats, democrat, bill clinton, immigration, immigrant, immigrants, elchapo, mexico, mexicans, latino,

Tabela 4. N-gramas Chave e Alvo

cujos n-gramas alvo favoráveis são bem definidos, como pode ser verificado na Tabela 4 para os alvos Hillary, Trump e Feminismo. Outra razão é a tendência à crítica através de comentários predominantemente negativos, como já discutido na Seção 4.2.

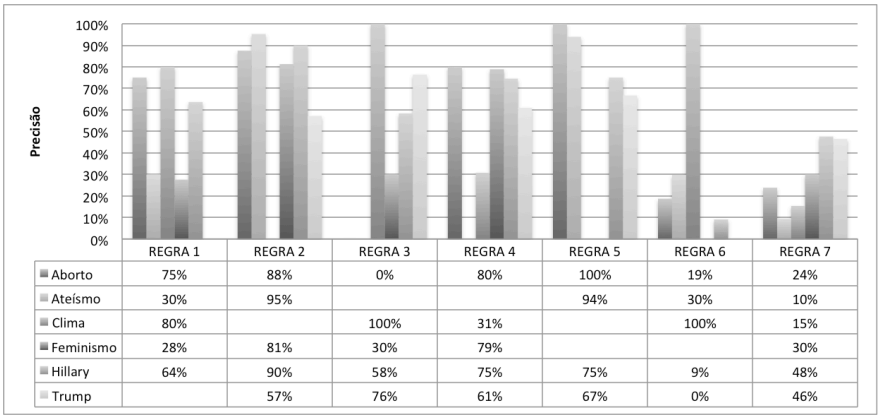
A regra 2 (CHAVE-CONTRA) também se aplica a um grande número de instâncias em quase todos os conjuntos de dados, à exceção do alvo *Clima*, para o qual nenhum n-grama deste tipo foi identificado. Note-se que esta regra é particularmente expressiva para os alvos *Feminismo* e *Trump*, apesar dos poucos n-gramas chave contrários identificados. Este é um comportamento importante, pois apenas a presença do n-grama já caracteriza o posicionamento do autor no tweet, tornando a detecção de posicionamento mais precisa. Contudo, a regra 1 oposta (CHAVE-FAVOR) só foi representativa no alvo *Feminismo*. Este fato pode ser explicado pela expressão de posicionamento através de críticas, frequente nestes datasets de uma forma geral.

A regra 7 (NEUTRO) foi representativa em todos os conjuntos de dados. Esta é a regra mais aplicada nos tweets relativos ao alvo *Clima*, onde há um elevado número de tweets do polaridade neutra (18,79%, contra um máximo de 5% nos demais conjuntos).

As regra 5 (ALVO-CONTRA-POSITIVO) e 6 (ALVO-CONTRA-NEGATIVO), foram aplicadas a um número elevado de instâncias do alvo *Ateísmo*, pois os tweets deste domínio tendem a referenciar alvos contrários ao alvo principal como os n-gramas "jesus"e "god". Ainda, sendo o ateístmo o alvo que apresenta o maior percentual de tweets positivos (60,03%), conforme a Tabela 2, houve uma boa representatividade de tweets filtrados pela regra 5. Já a regra 3 (ALVO-FAVOR-POSITIVO) foi representativa apenas para os alvos políticos (Hillary e Trump).

O gráfico da Figura 2 permite fazer uma análise do desempenho das regras no tocante à precisão do respectivo rótulo, considerando cada conjunto de dados. A precisão foi calculada considerando apenas o conjunto de teste.

As regras 1 e 2 são baseadas em n-gramas chave, sendo que a regra 2 (CHAVE-CONTRA) tem o desempenho mais consistente. Os resultados para a regra 1 (CHAVE-FAVOR) são inferiores, e inaceitáveis para dois alvos: *Feminismo* (28%) e *Ateísmo* (30%). A precisão para estas regras confirma a intuição de que o uso de certas expressões, em



	REGRA 1	REGRA 2	REGRA 3	REGRA 4	REGRA 5	REGRA 6	REGRA 7
Aborto	75%	88%	0%	80%	100%	19%	24%
Ateísmo	30%	95%			94%	30%	10%
Clima	80%		100%	31%		100%	15%
Feminismo	28%	81%	30%	79%			30%
Hillary	64%	90%	58%	75%	75%	9%	48%
Trump		57%	76%	61%	67%	0%	46%

Figura 2. Precisão x Regras x Alvos

particular hashtags, são uma forma mais direta de expressar um posicionamento, onde os resultados inferiores para a posição *Favorável* são explicados pela identificação de um número menor de n-gramas chave desta categoria (Tabela 4).

Já as regras 3 a 7 são dependentes tanto dos n-gramas alvo e das premissas sobre estes, quanto do desempenho da solução utilizada para identificação do sentimento. A medida F bastante baixa para a polaridade neutra é em grande parte responsável pelo resultado da regra 7 (NEUTRO), ruim para todos os conjuntos de dados, não sendo possível coletar evidências de que tweets sem posicionamento correspondem a textos sem sentimento. Este fraco desempenho teve forte impacto na medida F da abordagem proposta para detecção de posicionamento (Tabela 5), como será discutido na Seção 4.5. As regras para detecção de posicionamento *Contrário* baseadas em sentimento sobre um alvo apresentaram resultado superior, quando comparadas àquelas para o posicionamento *Favorável*.

Para o posicionamento *Favorável*, a regra 3 (ALVO-FAVOR-POSITIVO) apresenta resultado satisfatório apenas para 3 dos 5 conjuntos de dados aos quais foi aplicada. A variação nos resultados referentes a esta regra decorre do fato de ela ser pouco aplicada nos conjuntos de teste (no máximo 7% das instâncias). Ela também pode ter sido afetada pelo desempenho da detecção de sentimento positivo. Já o desempenho para a regra 6 (ALVO-CONTRA-NEGATIVO) é bom para um único caso: o alvo *Clima*, com precisão de 100%. Os resultados destas duas regras para o alvo *Clima*, em particular, podem ser parcialmente explicados pela predominância de tweets favoráveis neste conjunto de dados, comportamento não observado nos demais. Logo, não há evidências concretas de que a premissa subjacente à regra 6 seja válida como forma de expressão de endosso.

Quanto ao posicionamento *Contrário*, com exceção de um único caso, as regras 4 (ALVO-FAVOR-NEGATIVO) e 5 (ALVO-CONTRA-POSITIVO) apresentaram um bom resultado quando aplicadas (precisão mínima de 61% e 67%, respectivamente). No caso da regra 5, observa-se a relação com a identificação dos n-gramas alvo contrários, particularmente para *Aborto* e *Ateismo*, onde a precisão atingida foi de 100% e 94%, respectivamente. O alvo *Clima* foi a única exceção na regra 4, com apenas 31%, talvez explicado pela baixa representatividade deste posicionamento neste conjunto de dados. Mas, de uma forma geral, há evidências da contribuição destas duas regras para detecção de posicionamento.

Alvo	Classe	Rotulação Automática				Modelo Preditivo				Combinada				
		#Inst	Precisão	Revocação	Medida-F	#Inst	Precisão	Revocação	Medida-F	#Inst	Precisão	Revocação	Medida-F	SemEval
Aborto	**Contrário**	118	86,41	75,42	80,54	71	54,47	94,37	69,07	189	69,03	82,54	75,18	50,92
	Favorável	19	27,27	31,58	29,27	27	57,14	14,81	23,53	46	34,48	21,74	26,67	
	Sem Posic.	9	23,81	55,56	33,33	36	0,00	0,00	0,00	45	20,00	11,11	14,29	
	Med. Pond.	146	74,85	68,49	70,96	134	40,38	52,99	41,34	280	55,47	61,07	57,42	
Ateísmo	**Contrário**	108	94,44	47,22	62,96	52	59,18	55,77	57,43	160	77,67	50,00	60,84	49,58
	Favorável	21	29,82	80,95	43,59	11	19,35	54,55	28,57	32	26,14	71,88	38,33	
	Sem Posic.	3	9,52	66,67	16,67	25	62,50	20,00	30,30	28	24,14	25,00	24,56	
	Med. Pond.	132	82,23	53,03	58,83	88	55,15	45,45	46,11	220	63,36	50,00	52,95	
Clima	**Contrário**	5	30,77	80,00	44,44	6	11,11	16,67	13,33	11	22,73	45,45	30,30	42,09
	Favorável	50	95,00	38,00	54,29	73	66,00	45,21	53,66	123	74,29	42,28	53,89	
	Sem Posic.	4	15,38	100,00	26,67	31	29,41	48,39	36,59	35	24,68	54,29	33,93	
	Med. Pond.	59	84,16	45,76	51,58	110	52,69	44,55	46,65	169	60,66	44,97	48,22	
Feminismo	**Contrário**	103	79,31	66,99	72,63	80	60,49	61,25	60,87	183	70,24	64,48	67,24	45,79
	Favorável	29	28,21	37,93	32,35	29	16,67	10,34	12,77	58	24,56	24,14	24,35	
	Sem Posic.	4	30,00	75,00	42,86	40	36,00	45,00	40,00	44	35,00	47,73	40,38	
	Med. Pond.	136	66,96	61,03	63,17	149	45,39	46,98	45,90	285	55,50	53,68	54,36	
Hillary	**Contrário**	110	81,65	80,91	81,28	62	46,94	74,19	57,50	172	65,22	78,49	71,24	55,42
	Favorável	24	44,12	62,50	51,72	21	22,73	23,81	23,26	45	35,71	44,44	39,60	
	Sem Posic.	30	47,62	33,33	39,22	48	27,27	6,25	10,17	78	40,63	16,67	23,64	
	Med. Pond.	164	69,93	69,51	69,26	131	35,85	41,22	34,67	295	54,21	56,95	53,83	
Trump	**Contrário**	147	60,38	65,31	62,75	152	38,36	76,97	51,20	299	45,91	71,24	55,83	44,70
	Favorável	105	44,68	40,00	42,21	43	13,16	11,63	12,35	148	35,61	31,76	33,57	
	Sem Posic.	72	46,48	45,83	46,15	188	30,00	6,38	10,53	260	40,54	17,31	24,26	
	Med. Pond.	324	52,20	52,78	52,40	383	31,43	34,99	26,87	707	41,78	43,14	39,56	

Tabela 5. Resultados por Etapa e Alvo

4.5. Analise da Classificação do Posicionamento

A Tabela 5 mostra os resultados de precisão, revocação, medida F e a métrica do SEMEVAL por alvo e por classe, além das médias ponderadas das métricas divididos pelas etapas do processo: a) apenas para rotulação automática, b) apenas para a aplicação do modelo preditivo, e c) para a combinação das duas formas de rotulação. Nas colunas rotuladas **#Inst** estão contabilizados os tweets rotulados por cada método. Assim como na análise de precisão das regras, apenas os dados de teste foram usados para o cálculo destas métricas, mas o modelo preditivo foi criado com base no conjunto de dados de treino.

Considerando a medida F ponderada, é nítido o melhor desempenho obtido pela rotulação automática comparada aos modelos preditivos. A diferença é pequena apenas no alvo *Clima*, de apenas 5 pontos percentuais a favor da rotulação automática. Nos demais casos, a diferença varia de 12 até 30 pontos percentuais. Apesar da rotulação automática ser mais eficiente, ela não é uma solução abrangente, pois um elevado número de tweets não são filtrados pelas regras. O número de tweets filtrados por regras varia de 38% no conjunto de dados *Clima*, até um máximo de 66% no conjunto de dados *Trump*. Logo, o modelo preditivo atua como uma solução complementar à rotulação automática.

Por outro lado, o resultado do modelo preditivo é relacionado ao número e correção de instâncias rotuladas utilizadas no aprendizado. Como discutido na seção anterior, as regras são mais precisas para a classe *Contrária*, e apresentam resultado inaceitável para a classe neutra. Consequentemente, os resultados de previsão dos diferentes modelos para estas duas classes são os melhores e piores, respectivamente, à exceção do conjunto de dados *Clima*, devido às suas peculiaridades anteriormente citadas. Com raras exceções, a revocação é melhor que a precisão para a classe *Contrária*, devido ao maior número de instâncias usadas para treino. Na classe *Favorável*, o inverso ocorre.

Em geral, a medida F resultante desta técnica, para cada classe, é explicada por três fatores: a) o número/percentual de instâncias daquela classe usadas para treinamento, b) a precisão da rotulação automática das instâncias que compõem o corpus de treino e c) o número restante de instâncias para aplicação do modelo preditivo. A combinação de uma grande diferença no percentual de instâncias rotuladas e a baixa precisão (veja Tabela 5) justifica o resultado para as classes *Favorável* e *Sem posicionamento* na maioria dos alvos. Com relação ao último fator, pode-se ver claramente o seu impacto na classe *Contrária* para o alvo *Clima*.

Ao analisarmos por classe as métricas apresentadas na Tabela 5, fica evidente um melhor desempenho na detecção do posicionamento *Contrário* ao alvo principal, tanto na etapa de rotulação automática, quanto na etapa de aplicação do modelo preditivo. Considerando o resultado combinado, a medida F para a classe *Contrária* varia de 75,18% para o alvo *Aborto* a 55,84% para o alvo *Trump*. Esta afirmação desconsidera o conjunto de dados *Clima*, por suas características diferenciadas, que fazem com que o resultado para a classe *Favorável* seja o melhor, tanto na rotulação automática (54,29% de medida F), quanto no modelo preditivo (53,66%). Com exceção do alvo *Feminismo*, o pior resultado foi observado para a classe neutra.

Finalmente, a Tabela 5 mostra também o desempenho da abordagem para cada alvo, calculado de acordo com o mesmo critério do SEMEVAL, a saber, média da medida F considerando apenas as classes *Favorável* e *Contrária* (i.e. desprezando a classe *Sem posicionamento*). O vencedor da competição da tarefa não supervisionada, que envolveu apenas o conjunto de dados *Trump*, obteve um score de 56,28%. O segundo colocado atingiu 44,66%, enquanto nosso time obteve um resultado de 42,32%. Com o processo modificado para o uso combinado, tal como o descrito na Seção 3.2, atingimos um resultado de 44,7%, similar ao do segundo colocado.

Utilizando o resultado do SEMEVAL como baseline, é possível verificar que nossos resultados para o sistema completo são próximos do resultado vencedor da competição na tarefa semi-supervisionada, e superiores aos do segundo colocado para todos os conjuntos de dados, exceto para *Clima*. É importante ressaltar o resultado para o alvo *Hillary*, muito próximo do vencedor (55,42%) da competição não supervisionada. Ao comparar nossos resultados com os do vencedor da tarefa supervisionada, superamos este no alvo *Clima* por um ponto percentual e ficamos a apenas dois pontos no alvo *Hillary*.

5. Conclusões

Este trabalho discutiu uma abordagem semi-supervisionada para detecção de posicionamento em tweets, analisando seus componentes e o resultado final para um conjunto de seis diferentes alvos. A abordagem é simples, e apresenta um bom desempenho para alvos para os quais o posicionamento é expresso de variadas formas. Mas o melhor resultado conhecido para abordagem não-supervisionada, i.e. o vencedor do SEMEVAL 2016 com escore 56,28%, mostra que ainda há muito a evoluir.

O ponto chave de nossa abordagem são as regras, cuja representatividade e precisão foi avaliada. Por um lado, o aumento da precisão é em parte dependente da solução de identificação de sentimento, que ainda é falha. Por outro lado, algumas regras ainda carecem de validação como forma de expressão de posicionamento.

Como trabalhos futuros, devemos verificar a possibilidade de criação de novas regras que rotulem tweets que estão sendo descartados pela abordagem atual. Para isso podem ser definidas regras com base em quantidades de menções aos n-gramas, desta forma, passaríamos a rotular tweets com menções a alvos dos dois lados do posicionamento, aumentando sua cobertura. O procedimento de detecção de polaridade dos tweets deve ser alvo de aprimoramento, pois influencia muito os resultados. Para obter melhores resultados na etapa de aplicação do modelo preditivo, podem ser estudadas melhorias como a utilização de outros algoritmos de classificação ou o uso de comitês de algoritmos. Podemos ainda explorar os atributos utilizados para a construção do classificador, acrescentando atributos relacionados a sentimento ou bi-gramas e tri-gramas. Pode-se ainda estudar outras formas para a identificação dos n-gramas chave e alvo, tais como

métodos de extração de tópicos, ou explorar as relações estáticas e dinâmicas do Twitter entre perfis notadamente apoiadores ou opositores ao alvo principal.

Agradecimentos

Este trabalho é parcialmente financiado pelo CNPq (Projeto 459322/2014).

Referências

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., e Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. Em *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pgs. 1–9. Association for Computational Linguistics.

Dias, M. e Becker, K. (2016). INF-UFRGS-OPINION-MINING: Automatic generation of a training corpus for unsupervised identification of stance in tweets. Em *Proceedings of the International Workshop on Semantic Evaluation - Task 6*, SemEval '16, San Diego, CA, USA. To appear.

Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. Em *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014, Pensacola Beach, Florida, May 21-23, 2014.*

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., e Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. Em *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Mohammad, S. M., Zhu, X., Kiritchenko, S., e Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Rajadesingan, A. e Liu, H. (2014). Identifying users with opposing opinions in twitter debates. Em *Social Computing, Behavioral-Cultural Modeling and Prediction*, pgs. 153–160. Springer.

Somasundaran, S. e Wiebe, J. (2009). Recognizing stances in online debates. Em *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pgs. 226–234. Association for Computational Linguistics.

Tang, J., Nobata, C., Dong, A., Chang, Y., e Liu, H. (2015). Propagation-based sentiment analysis for microblogging data. Em *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pgs. 577–585.

Thomas, M., Pang, B., e Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. Em *Proceedings of the 2006 conference on empirical methods in natural language processing*, pgs. 327–335. Association for Computational Linguistics.