Bazinga! Caracterizando e Detectando Sarcasmo e Ironia no Twitter

Pollyanna Gonçalves[1], Daniel Hasan Dalip[1], Julio C. S. Reis[1], Johnnatan Messias[1], Filipe Ribeiro[2], Philipe Melo[1], Leandro Araújo[1], Fabrício Benevenuto[1], Marcos Gonçalves[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Brasil

[2]Universidade Federal de Ouro Preto (UFOP) – Brasil

{pollyannaog,hasan,julio.reis,johnnatan,philipe}@dcc.ufmg.br

{leandroaraujo,fabricio,mgoncalv}@dcc.ufmg.br

{filipe}@decea.ufop.br

Abstract. *Sarcasm and irony are widely used forms of speech used inside and outside the Web, having the power to transform a sentence regarding its polarity or sense. The ability of characterizing and detecting sarcastic and ironic messages on data collected from Web could improve many decision-making systems based on Natural Language Processing (NLP) such as the sentiment analysis, text summarization and review ranking systems. In this work, we propose some approaches to the task of characterization and detection of sarcasm and irony in messages posted on Twitter online social network. Using an automatically collected dataset with the hashtags "#sarcasm" and"#irony", and by exploiting a large set of characterization and classification techniques, our results show satisfactory rates of accuracy and Macro-F1.*

Resumo. *Sarcasmo e ironia são formas de discurso muito utilizadas dentro e fora da Web, tendo o poder de transformar características como polaridade ou sentido de uma sentença. Ser capaz de caracterizar e detectar mensagens sarcásticas ou irônicas em dados coletados da Web pode aprimorar diversos sistemas de tomada de decisão baseados em Processamento de Linguagem Natural (PLN) como a tarefa de análise de sentimentos, sumarização de textos e sistemas de ranqueamento de reviews. Nesse trabalho, propomos diversas abordagens para a caracterização e posterior classificação de sarcasmo e ironia em mensagens postadas na rede social online Twitter. Utilizando uma base automaticamente coletada de tweets com as hashtags "#sarcasm"e "#irony", e usando uma larga gama de técnicas de caracterização e classificação, nossos resultados de detecção alcançaram taxas satisfatórias de acurácia e Macro-F1.*

1. Introdução

A habilidade de identificar sarcasmo e ironia em diálogos informais publicados na Web vem recebendo a atenção de diversas pesquisas científicas na computação e em várias outras áreas do conhecimento, como por exemplo, na linguística [Cheang and Pell 2011] e na sociologia [Ball 1965]. No mundo real, as pessoas são capazes de compreender tais características em uma conversa devido a vários fatores contextuais tais como os

gestos do locutor e seu tom de voz [Gibbs and Colston 2007]. No entanto, devido a essa natureza intrínseca de interpretação do ser humano, a tarefa de automaticamente detectar sarcasmo e ironia se torna difícil em dados provenientes da Web. Uma das principais características desse tipo de mensagem é a capacidade de desempenhar um papel de inversor de polaridade ou até mesmo de seu sentido, utilizando-se de várias técnicas linguísticas tais como simples jogos de palavras. Reconhecer essas mudanças é um aspecto extremamente importante na melhora da performance de algoritmos propostos na mineração de opinião, como por exemplo para a detecção de sentimentos em dados da Web [Hu and Liu 2004, Popescu and Etzioni 2005]. Além da pesquisa científica, detecção de sarcasmo e ironia em dados da Web vêm atraindo esforços em diversos setores. Recentemente, a agência do Serviço Secreto dos Estados Unidos anunciou a contratação de desenvolvedores para a construção de um detector de sarcasmo voltado para redes sociais online [BBC]. Essa notícia demonstra a relevância do tema na atualidade e ressalta, principalmente, a complexidade existente por trás dessa tarefa.

Redes sociais online se tornaram um meio propício para a coleta de dados por pesquisadores e companhias em larga escala [Cha et al. 2010]. Nesses ambientes, usuários expressam suas opiniões, fatos de seu cotidiano e informações de situações que presenciam em tempo real [Cha et al. 2012]. Como resultado, um grande número de estudos utiliza-se desses ambientes para o monitoramento de assuntos do momento, memes[1], de eventos em escala global [Gomide et al. 2011, Lamb et al. 2013] e desastres [Sakaki et al. 2010]. Para exemplificar o grande uso de redes sociais nos dias de hoje, o Twitter, atualmente considerado como uma das redes sociais online mais populares da Web, conta com mais de 280 milhões de usuários ativos por mês e 500 milhões de tweets[2] postados diariamente.

Aqui, damos um importante passo em direção a caracterização e detecção de sarcasmo e ironia no Twitter. Nessa rede social, mensagens podem ser associadas a *hashtags* como "#NowPlaying" (utilizado por usuários para indicar a música sendo ouvida naquele momento) e "#FollowFriday" (utilizado para recomendar seguidores na rede). Utilizamos essa propriedade para construir uma base de dados rotulada com cerca de 2.628 tweets contendo sarcasmo (aqueles que contêm a hashtag #Sarcasm), 2.628 tweets contendo ironia (associados à hashtag #Irony) e 2.628 tweets com contextos aleatórios. Nosso processo de caracterização envolve o uso do software *Linguistic Inquiry and Word Count* (LIWC) como uma técnica de Processamento de Linguagem Natural (PLN), análise de estruturas gramaticais, técnicas de aprendizado de máquina que exploram o texto (*bag-of-words*) da mensagem, análise de sentimento e detecção de atributos de usuários para a classificação de mensagens como sendo ou não sarcásticas ou irônicas.

O restante desse artigo está organizado da seguinte forma. A Seção 2 descreve os trabalhos relacionados. Em seguida, na Seção 3, apresentamos a metodologia utilizada para a caracterização e detecção de sarcasmo e ironia no Twitter. A análise e discussão dos resultados são apresentados na Seção 4. Por fim, a Seção 5 apresenta conclusões e direções para trabalhos futuros.

[1]Termo que descreve um conceito que se tornou "viral", ou seja, aquele que se espalhou rapidamente pela Internet.

[2]Mensagem com um máximo de 140 caracteres postadas no Twitter.

2. Terminologia e Conceitos

Nesse trabalho, propomos uma caracterização e detecção de sarcasmo e ironia em dados coletados do Twitter. Uma das principais dificuldades presentes na tarefa de detectar tais tipos de mensagens na Web está associada a falta de acordo entre a maioria dos pesquisadores (sociólogos, psicólogos, cientistas da computação, etc.) em como definir ou diferenciar sarcasmo de ironia. Enquanto várias pesquisas propõem que sarcasmo e ironia são termos associados a um mesmo fenômeno linguístico [Kreuz and Glucksberg 1989a], há estudos focados em mostrar evidências que os diferenciam [Singh 2012].

Considerando que nossa base possui mensagens de usuários que diferenciam sarcasmo de ironia com o uso das hashtags "#sarcasm" e "#irony", propomos estratégias para caracterização e detecção de mensagens sarcásticas e irônicas separadamente. Outro fator importante que consideramos para separação das bases com sarcasmo e ironia foi a baixa quantidade de mensagens que contêm as hashtags "#sarcasm" e "#irony" ao mesmo tempo (cerca de 0,0007% da base).

Para melhor entendimento da diferenciação entre sarcasmo e ironia, definimos brevemente abaixo os conceitos desses dois temos:

- **Sarcasmo:** No sarcasmo, há um uso de instrumentos linguísticos indiretos para a ridicularização ou zombaria, muitas vezes considerado grosseiros e ofensivos, sendo utilizados para fins destrutivos [Singh 2012]. Um exemplo de sarcasmo pode ser visto no tweet "*Super glad I got a sinus infection during finals week! #sarcasm*", em que é ontrastado um sentimento de agradecimento com o surgimento de um contratempo negativo em um período possivelmente conturbado.
- **Ironia:** A ironia pode ser considerada como uma discordância, ou incongruência, entre o que se diz e o que se entende, ou do que se espera e do que realmente ocorre [Singh 2012]. Esse tipo de mensagem geralmente vêm acompanhado de um tom de brincadeira e possui menor peso ofensivo do que o sarcasmo. Um exemplo de ironia pode ser notado no tweet "*Steve Jobs did not allow his kids to use iPads #irony*", onde, em um tom engraçado o usuário encontra uma possível incongruência de uma situação.

A seguir, apresentamos os trabalhos relacionados e posicionamos nosso trabalho em relação à literatura.

3. Trabalhos Relacionados

Esforços na caracterização e detecção sarcasmo e ironia vêm sendo abordado pela literatura em diferentes áreas de pesquisa, tais como linguística [Cheang and Pell 2011] e psicologia [Kreuz and Glucksberg 1989b, Gibbs and Colston 2007]. No entanto, apesar dos diversos estudos em diferentes área, ainda há desafios a serem resolvidos, princialmente nas áreas de Ciência da Computação, como a Mineração de Opinião, onde o objetivo é detectar tais expressões de forma automática.

Estudos recentes mostram, por exemplo, que é possível que sentenças sarcásticas sejam construídas utilizando-se padrões e fórmulas, os quais facilitam sua identificação [Kreuz and Caucci 2007]. Diante desse fato, surge a possibilidade da criação de métodos que busquem por fatores gramaticais, como advérbios e interjeições comuns em sentenças classificadas como sarcásticas ou irônicas, e que os utilizem na

classificação de mensagens desse tipo na Web. Outros evidenciam como a presença de ironia em mensagens pode dificultar a classificação de sentimentos em mensagens da Web [Carvalho et al. 2009].

Mais relacionados ao nosso estudo, alguns trabalhos utilizam a técnica de filtragem por hashtags como uma estratégia para a identificação de mensagens em redes socias contendo sarcasmo ou ironia. Gonzales-Ibáñes et. al [González-Ibáñez et al. 2011] construíram um corpus de mensagens com teor sarcástico postadas no Twitter, utilizando como rótulo para a coleta as hashtags "#sarcasm" e "#sarcastic". Esse estudo esteve focado na diferenciação de tweets contendo sarcasmo de tweets com teor positivo ou negativo e posterior identificação de sarcasmo na rede social online. O método desenvolvido com o uso de técnicas de aprendizado de máquina se baseou no uso de atributos léxicos como o LIWC e o WordNet Affect (WNA) [Valitutti 2004], atributos chamados pragmáticos, como emoticons positivos e negativos e tweets destinados a um usuário específico. Resultados mostraram que atributos léxicos apenas não são suficientes para distinguir mensagens com sarcasmo de mensagens positivas ou negativas, ao contrário dos atributos pragmáticos. O estudo também ressaltou as dificuldades na identificação de mensagens desse tipo, comparando resultados da classificação humana com resultados do método de aprendizado de máquina proposto.

Outro estudo, de Liebrecht et al. [Liebrecht et al. 2013], realizou a coleta de milhares de tweets no idioma Holandês contendo hashtags associadas ao sarcasmo e desenvolveram um método baseado na técnica de aprendizado de máquina *Balanced Winnow* [Littlestone 1988] para alcançar uma performance de 75% de acurácia na detecção de mensagens desse tipo. O algoritmo utilizado é considerado estado da arte na tarefa de classificação de textos e é capaz de associar pesos as classes com o objetivo de, por exemplo, inspecionar atributos com maior ganho de informação em uma determinada classe. Já Li et al. [Li et al.] utilizaram hashtags como *baseline* para a tarefa de recuperação de informação e classificação para a construção de um corpus de declarações relevantes para a detecção de sarcasmo.

Diferente dos trabalhos abordados na literatura, nosso estudo faz o uso abordagens de algoritmos de aprendizado de máquina que se baseiam em *Bag-of-words*, *features* associadas a componentes cognitivos, emocionais e estruturas gramaticais para a classificação das mensagens, além de caracterizá-las em relação as suas características linguísticas, atributos dos usuários que as postaram e dos próprios tweets compartilhados e em relação a polaridade (positivo, negativo ou neutro) dos tweets. Além disso, comparamos o uso de três estratégias de classificação e mostramos qual das estratégias tem a capacidade de alcançar as melhores taxas de acurácia e Macro-F1.

Na próxima seção, discutimos a metodologia utilizada nesse trabalho. Descrevemos o processo de coleta da base de dados utilizada, definimos os algoritmos e técnicas propostas para a tarefa de caracterização e detecção de sarcasmo e ironia no Twitter, assim como as métricas utilizadas para a avaliação das abordagens consideradas.

4. Metodologia

Nesta seção, descrevemos nosso processo para a coleta de dados e as métricas utilizadas para a avaliação da nossas abordagens de detecção de sarcasmo e ironia no Twitter.

4.1. Coleta da Base de Dados

Nesse trabalho, utilizamos a API do Twitter[3] para coletar mensagens a serem detectadas como sarcásticas ou irônicas postadas por usuários da rede social. O uso da API restringe a coleta a apenas 1% dos tweets públicos compartilhados na rede. Apesar da baixa permissão para coleta, 1% do total de tweets publicados diariamente significa milhares de mensagens aleatórias possíveis de serem coletadas por dia. Coletamos tweets públicos no período de Fevereiro de 2014 a Fevereiro de 2015. Ao final desse processo, contamos com 7.884 mensagens de onde conseguimos extrair 2.628 tweets contendo as hashtags "#sarcasm", 2.628 tweets contendo "#irony" e 2.628 contendo tweets com contexto aleatório. Além da mensagem postada, cada tweet possui informações como *timestamp*, indicador de *retweet* e favoritos, além de informações sobre o usuário que postou o conteúdo, como ID, nome de usuário e número de seguidores e seguidos.

4.2. Caracterização

Com o intuito de investigar e encontrar características próprias de mensagens contendo sarcasmo ou ironia no Twitter, damos um passo na caracterização da nossa base de dados. Nesse trabalho, propomos uma caracterização dos tweets coletados baseando-se em 3 estratégias: (i) Características linguísticas do conteúdo; (ii) Atributos de usuários e tweets; e (iii) Análise de sentimentos. A caracterização por características linguísticas do conteúdo tem como objetivo um maior entendimento do uso de atributos linguísticos (ex.: estrutura e gramática das mensagens) em mensagens definidas como contendo sarcasmo ou ironia. Nesse trabalho, utilizamos como estratégia para essa caracterização a ferramenta *Linguistic Inquiry and Word Count* (LIWC) [Tausczik and Pennebaker 2010]. O uso dessa ferramenta é importante para nosso estudo pois ele nos permite estimar componentes emocionais, cognitivos, estruturais e gramaticais de um texto fornecido como entrada, baseado na utilização de dicionários contendo palavras e categorias associadas a cada uma. O software está disponível em `http://www.liwc.net/`.

A caracterização focada em atributos de usuários e tweets nos permitirá analisar as diferenças do comportamento dos usuários nas redes e dos tweets postados (ex.: quantidade de seguidores e quantidades de amigos, quantidades de retweets favoritos, menções a outros usuários, URLs, mídias e hashtags) em mensagens desse tipo.

Por fim, a análise de sentimentos nos apresentará como essas mensagens se diferem em termos de polaridade (positivo, negativo ou neutro).

4.3. Predição de Sarcasmo e Ironia

Para o aprendizado e detecção de sarcasmo ou ironia nos tweets coletados, utilizamos o método de aprendizado de máquina supervisionado *Support Vector Machine* (SVM) [Tsochantaridis et al. 2005]. Nessa técnica, cada tweet T foi representado por um vetor de atributos $T = (f_{1t}, f_{2t}...f_{nt})$ e uma classe alvo c. Essa classe alvo, no caso da identificação de ironia, será "ironia" caso seja ironia e "não-ironia" caso não seja ironia. No caso da identificação de sarcasmo, também será uma classificação binária onde "sarcasmo" quando a mensagem possui sarcasmo e "não-sarcasmo" caso contrário.

Este vetor de atributos do tweet é composto por indícios que poderiam inferir se um determinado tweet contem sarcasmo/ironia ou não. Com essa estratégia, podemos

[3] `https://dev.twitter.com/`

representar um único tweet de 3 formas: (i) Utilizando as palavras do tweet (técnica *Bag-of-Words* (BOW)); (ii) Utilizando os atributos extraídos do LIWC; ou (iii) Utilizando ambas as representações. Essas três representações serão chamadas nesse trabalho de BOW, LIWC e BOW+LIWC, respectivamente.

Na representação BOW, cada tweet é representado por um vetor $T = (w_{1t}, w_{2t}, w_{3t}, ..., w_{vt})$. Onde v é o tamanho do vocabulário e w_{it} é o peso da palavra i no tweet t, que queremos prever. O peso w_{it} é calculado através da métrica TF-IDF [Baeza-Yates and Ribeiro-Neto 2011] de cada palavra, como na equação abaixo:

$$w_{it} = TF_{it} \times IDF_i = (1 + log_2(f_{it})) \times log_2(\frac{N}{n_i}) \quad (1)$$

onde f_{it} é a frequência da palavra i no tweet t, N é a quantidade total de tweets na coleção e n_i é a quantidade de tweets que a palavra i aparece.

A representação LIWC utiliza suas 85 categorias (tratadas aqui como atributos) para representar um tweet. Por fim, a representação BOW+LIWC utiliza a concatenação dos pesos por palavras (Eq. 1) e os 85 atributos da representação LIWC.

Para medir o desempenho da abordagem de detecção proposta, utilizamos as medidas convencionais de Recuperação de Informação: revocação, precisão, acurácia e a macro F1. A revocação (R) de uma classe é a fração do número de mensagens corretamente classificadas pelo número de mensagens nessa classe. Precisão (P) da classe é a fração do número de mensagens corretamente classificadas pelo total de mensagens preditas como mensagens dessa classe. Com o objetivo de explicar tais métricas, construímos a matriz de confusão apresentada no quadro a seguir:

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

Cada posição neste quadro representa o número de elementos em cada classe, e como essas foram previstas pelo modelo utilizado na classificação. Assim, conforme essa matriz, a precisão (P_{pos}) e revocação (R_{pos}) para a classe positiva podem ser calculadas da seguinte forma: $P_{pos} = a/(a + b)$ e $R_{pos} = a/(a + c)$. Já a acurácia e macro F1, podem ser calculados como: $A = (a + d)/(a + b + c + d)$ e $F1 = 2 \cdot (P \cdot R)/(P + R)$, respectivamente.

Em nossos experimentos utilizamos um procedimento de validação cruzada em cinco partições (*folds*) em que quatro partições são usadas para treino e um para teste e o procedimento é repetido cinco vezes, com variação das partições utilizando uma estratégia *round-robin*. Resultados correspondem à média das cinco partições de teste. Parâmetros do classificador SVM foram descobertos utilizando a ferramenta *grid* disponível no pacote LibSVM: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, utilizando validação cruzada dentro do conjunto de treino. Por estarmos trabalhando com classificação de texto com alta dimensionalidade e esparsidade, utilizamos um SVM com *kernel* linear.

5. Resultados

5.1. Caracterização

Nesta seção, apresentamos os resultados da etapa de caracterização de sarcasmo e ironia nos dados coletados do Twitter.

5.1.1. Características Linguísticas do Conteúdo

Um dos nossos objetivos neste trabalho é compreender características linguísticas de determinados conteúdos que contenham sarcasmo e ironia. Para isso, comparamos e contrastamos as amostras de dados contendo sarcasmo e ironia com a que não contém, conforme detalhamento apresentado em seções anteriores, com a utilização do LIWC [Tausczik and Pennebaker 2010].

O LIWC consiste em um conjunto de palavras agrupadas em categorias associadas a quatro classes gerais: *Linguistic Processes* (LP) (e.g., advérbios, pronomes), *Psychological Processes* (PP) (e.g., emoções negativas e positivas), *Personal Concerns* (PC) (e.g, trabalho), e *Spoken Categories* (SC). Além, disso a ferramenta nos fornece taxas de ocorrência de *Punctuation* (PT) (e.g, vírgula, exclamação).

Inicialmente, executamos a ferramenta e obtivemos o percentual de ocorrência de palavras associadas a cada uma das 85 categorias, para os todos os cenários (e.g. aleatório, sarcasmo e ironia). Considerando que nosso propósito é investigar características intrínsecas ao texto que contém sarcasmo e ironia, focamos nos dicionários ou categorias do LIWC que possuem percentual significativo em relação as mensagens que não contém sarcasmo e ironia. Por exemplo, para a categoria *3rd pers plural* o percentual de ocorrência para os dados aleatórios foi de 0,35 enquanto para os dados que contém ironia o percentual foi de 0,74. Neste caso, a razão entre os valores é de 2,11 (0,74/0,35), o que indica que a ocorrência de palavras da categoria *3rd pers plural* é 2,11 vezes maior em ironia (em relação aos dados aleatórios). Com base no exemplo apresentado, definimos como significativa a classe cuja razão é menor que 0,5 (ocorrência menor ou igual a metade) ou maior que 2 (ocorrência maior ou igual ao dobro). Acreditamos que isso nos forneça um indício mais preciso de categorias que realmente sejam úteis para a diferenciação dos conteúdos analisados neste trabalho. Os resultados obtidos para as classes significativas do LIWC são apresentados na Tabela 1.

Percebemos que o tweet contendo sarcasmo faz uso significativo de palavras que expressam concordância, ainda que de maneira mais informal, com uso de expressões como "*Er*", "*hm*"e "*umm*"(ver classe *Nonfluencies*, na Tabela 1). Outro aspecto observado está relacionado ao uso intenso da pontuação dois pontos (":"), que é utilizada para o anúncio ou introdução de um esclarecimento ou citação. Já o conteúdo irônico tem presença significativa de pronomes na terceira pessoa do plural. Estudos anteriores indicam que essas são características de uma linguagem mais informal [Tausczik and Pennebaker 2010].

Categoria	Sarcasmo /Aleatório	Ironia /Aleatório	Exemplo de Palavras	Classe LIWC
3rd pers plural	1,60	**2,11**	They, their, they'd	*LP*
Assent	**2,63**	1,25	Agree, OK, yes	*SC*
Nonfluencies	**3,00**	1,22	Er, hm, umm	
Colons	**0,37**	0,53	-	*PT*

Tabela 1. Análise das classes significantes do LIWC

	Aleatório	Sarcasmo	Ironia
Ocorrência de mídias	32,90%	12,60%	24,60%
Ocorrência de URL's	29,40%	8,90%	12,80%
Ocorrência de mídia ou URL's	43,10%	14,80%	24,5%

Tabela 2. Ocorrência de mídia e URL's em tweets nas 3 categorias

5.1.2. Atributos de Usuário

Como discutido anteriormente, gostaríamos de analisar as diferenças dos tweets e atributos de usuários no Twitter em mensagens postadas com conteúdo sarcástico ou irônico. Para isso, foi necessário realizar uma análise das categorias dos tweets (tweets com a hashtag de sarcasmo, tweets com a hashtag de ironia e tweets aleatórios), cada categoria possuindo no total 2.628 tweets.

Nesse processo de caracterização, podemos determinar algumas especificidades, apresentadas na Tabela 2. Primeiramente notamos que a ocorrência de mídias (fotos ou vídeos) nos tweets da base aleatória é cerca de 2.61 vezes mais frequente do que na base de sarcasmo e 33.7% mais frequente do que na base de ironia. O mesmo ocorre com urls (ou links) que também são mais frequentes nos tweets da base aleatória do que nas outras bases.

Também analisamos a ocorrência de mídia ou URL nos tweets. O objetivo desta análise é mostrar a presença de alguma informação externa, seja foto, vídeo, URL, etc. Uma possível razão para esta constatação pode ser consequência da própria permissão de postagem do Twitter, que é reduzida a um espaço de 140 caracteres. Uma URL, mesmo encurtada, por exemplo, ocupa um espaço essencial para a construção de uma mensagem de sarcasmo ou ironia, pois estas últimas dependem totalmente do acompanhamento de um texto para transmitirem seu significado. Outro motivo que poderia levar a esse comportamento, seria que ironia e sarcasmo são realizados sobre um acontecimento ou situação que é de "senso comum" ou então de conhecimento de um grupo de amigos. Portanto, para esses acontecimentos ou situações já não é necessária nenhuma informação externa (URL ou mídia) para que a ironia e sarcasmo façam sentido.

Ao analisarmos apenas os 10% tweets mais favoritados de cada base a tendência continua. Também pode-se notar que a ocorrência de informação externa entre os 10% tweets mais favoritados da base aleatória é ainda maior, alcançando 58.4%. Os valores apresentados confirmam, que mesmo os tweets de sarcasmo e ironia mais populares possuem menos mídia e URL's do que os tweets dos demais assuntos.

	Ironia	**Sarcasmo**
Positivo	9%	33%
Negativo	**43%**	**38%**
Neutro	16%	5%
Não-definido	32%	24%
Concordância	*74%*	*79%*

Tabela 3. Caracterização de sarcasmo e ironia com análise de sentimentos

5.2. Análise de Sentimentos

No contexto de análise de sentimentos (ou mineração de opinião) em dados coletados da Web, a tarefa de se detectar sarcasmo ou ironia se torna muito importante. Isso porque, como discutido em seções anteriores, tais estratégias linguísticas têm o poder de transformar o sentimento de uma sentença, convertendo-a de uma sentença positiva para negativa, ou vice-versa. Nesse trabalho, temos com um dos nossos objetivos mostrar como tweets com sarcasmo e ironia estão associados a uma polaridade (positivo, negativo e neutro). Para isso, solicitamos que 3 voluntários rotulassem 100 mensagens aleatórias da nossa base de dados coletada contendo a hashtag "#irony" e 100 com a hashtag "#sarcasm" como positiva, negativa, neutral e não-definida (quando não conseguiam associar à mensagem um dos sentimentos anteriores). É válido ressaltar que os entrevistados tinham ciência desses rótulos antes de começarem a tarefa. Cada voluntário fez sua análise de forma independente, sem a influência dos demais. A porcentagem de concordância entre eles foi de 74% nas mensagens com a hashtag de ironia e 79% nas mensagens com a hashtag de sarcasmo. O resultado da caracterização de mensagens com ironia e sarcasmo em relação a polaridade é apresentado na Tabela 3.

Como podemos perceber, mensagens associadas as hashtags de ironia e sarcasmo tendem a estar mais relacionadas a polaridade negativa. Também podemos notar um percentual significativo de mensagens com rótulo "não-definido", sugerindo que mensagens desse tipo são difíceis de serem identificadas até mesmo por seres humanos.

5.3. Identificação Automática de Sarcasmo e Ironia

Nesta seção é apresentado o resultado da identificação automática de sarcasmo e ironia utilizando as 3 representações e o método de aprendizado de máquina descritos anteriormente. As Tabelas 4 e 5 mostram resultados de acurácia, Macro-F1, precisão e revocação por classe para identificação de sarcasmo e ironia, respectivamente. Como discutido anteriormente, a precisão indica a razão entre o número de instâncias corretamente classificadas com uma determinada classe c e a quantidade de instâncias classificadas como c. A revocação é a razão entre o número de instâncias corretamente classificadas com uma determinada classe c e a quantidade de instâncias que são realmente desta classe.

Analisando a performance para identificação de sarcasmo (Tabela 4), podemos observar que o BOW conseguiu uma melhor performance e, ao utilizar apenas o LIWC, não foi possível identificar corretamente sarcasmo, prevendo quase todas as instâncias como não-sarcasmo (revocação da classe sarcasmo baixa e da classe não-sarcasmo muito alta). Por esse motivo, ao utilizarmos a representação BOW+LIWC, não foi possível conseguir uma melhoria.

Representação	Acurácia	Macro F1	Classe "sarcasmo"		Classe "não-sarcasmo"	
			Precisão	Revocação	Precisão	Revocação
BOW	71,60%	71,79%	70,17%	76,25%	74,02%	67,54%
LIWC	50,24%	36,90%	58,63%	5,52%	50,12%	94,91%
BOW+LIWC	71,15%	71,13%	70,29%	73,55%	72,26%	68,83%

Tabela 4. Precisão e revocação por classe ao prever se um tweet contém sarcasmo ou não

Representação	Acurácia	Macro F1	Classe "ironia"		Classe "não-ironia"	
			Precisão	Revocação	Precisão	Revocação
BOW	71,51%	71,50%	71,33%	72,00%	71,74%	71,05%
LIWC	67,46%	67,42%	66,43%	70,57%	68,64%	64,36%
BOW+LIWC	76,00%	75,99%	76,60%	74,92%	75,46%	77,09%

Tabela 5. Precisão e revocação por classe ao prever se um tweet contém ironia ou não

Na identificação de ironia, como podemos observar na Tabela 5, as representações BOW e LIWC conseguem uma performance similar e, ao utilizar BOW+LIWC, conseguimos uma melhoria na performance demonstrando que cada uma das representações possui informações complementares. Por exemplo, melhorias de até 6,3% podem ser observadas com a combinação da representação LIWC com BOW tanto em acurácia quanto em Macro-F1.

Comparando os resultados da identificação de ironia e de sarcasmo, percebemos uma dificuldade maior na identificação de sarcasmo, onde não foi possível identificar utilizando apenas a representação LIWC. A dificuldade em tratar mensagens contendo sarcasmo também esteve presente na tarefa de caracterização de mensagens desse tipo com o uso de análise de sentimentos e na identificação de atributos de usuário, como discutido anteriormente. Esses resultados mostram a complexidade nesse tipo de classificação, sugerindo uma possível necessidade de se combinar técnicas em um esforço de alcançar melhores resultados para mensagens com teor sarcástico.

6. Conclusões e Trabalhos Futuros

Sarcasmo e ironia são formas de discurso linguístico ricos e complexos que vêm atraindo a atenção de pesquisadores na área de mineração de dados da Web devido a sua dificuldade de detecção nesse ambiente. Nesse trabalho, damos um passo na tarefa de caracterização e detecção de sarcasmo no Twitter. Contamos com uma base de dados significativa coletada para fins desse trabalho, contendo 2.628 tweets com conteúdo aleatório, 2.628 tweets com as hashtags "#sarcasm" e 2.628 tweets com a hashtag "#irony". Propomos diferentes abordagens para a caracterização de sarcasmo e ironia e também utilizamos técnicas de aprendizado de máquina para a detecção desse tipo de mensagem no Twitter.

Nosso processo de caracterização foi resultante da análise de três estratégias, baseadas em características linguísticas do conteúdo, onde analisamos resultados provenientes da ferramenta LIWC, em atributos de usuários, onde analisamos as diferenças entre tweets com sarcasmo e ironia em relação ao comportamento dos usuários nas redes e dos tweets postados, e em análise de sentimentos, onde diferenciamos as duas categorias de mensagens em relação a polaridade (positivo, negativo ou neutro). Nossos resultados mostraram que mensagens com teor sarcástico ou irônico fazem o uso de palavras na terceira pessoa

do plural quase 2 vezes mais que mensagens aleatórias, além de desse tipo de mensagens utilizar mais expressões como "*Er*", "*hm*" e "*umm*".

Nossa caracterização em relação a atributos de usuários e tweets mostraram que mensagens com conteúdo irônico, e especialmente de sarcasmo, possuem menos informações externas do que os tweets comuns indicando que, em geral, o mais importante da ironia e sarcasmo é o conteúdo textual e um fundamentado conhecimento sobre o assunto abordado. Uma abordagem que possivelmente proporcionaria melhores resultados, mas que não foi considerada inicialmente seria a verificação do contexto dos tweets dos usuários. Informações como a quantidade de tweets de sarcasmo e ironia anteriormente feitos pelo usuário ou a frequência de postagem de sarcasmo e ironia de seus amigos podem auxiliar na caracterização de usuários irônicos ou sarcásticos. Além disso, também mostramos que tweets sarcásticos e irônicos são mais negativos que positivos. Essa análise foi possível com um processo de rotulação feita por três voluntários que rotularam mensagens com sarcasmo e ironia com relação a sua polaridade (positivo, negativo, neutro, e não-definido).

Por fim, apresentamos nossos resultados do processo de detecção de sarcasmo e ironia no Twitter, com o uso do classificador supervisionado SVM junto ao algoritmo para classificação de textos TF-IDF. Utilizamos técnicas de *Bag-of-Words* (BoW) e categorias do LIWC para a etapa de detecção. Nossa abordagem se deu em três estratégias: (i) Detecção apenas com o uso do BoW, alcançando resultados acima dos 70% para sarcasmo e ironia; (ii) Detecção utilizando apenas as categorias do LIWC como atributos, alcançando resultados em torno de 36% para sarcasmo e 67% para ironia; e (iii) Detecção utilizando BoW e categorias do LIWC, alcançando resultados acima dos 71% para sarcasmo e 76% para ironia. Comparando esses percebemos uma maior dificuldade na identificação de sarcasmo, assim como a conclusão que apenas os atributos do LIWC aplicados ao classificador SVM não é suficiente para a tarefa de detecção de sarcasmo e ironia no Twitter. Como trabalho futuro, seria interessante observarmos quais termos (e tipos de termos) da técnica *Bag-of-Words* são os mais importantes para conseguirmos identificar sarcasmo e ironia. Dessa forma, seria possível propor atributos mais representativos para a tarefa.

7. Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e a Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG).

Referências

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Ball, D. W. (1965). Sarcasm as sociation: The rhetoric of interaction. pages 190–198.

BBC. Us secret service seeks twitter sarcasm detector. `http://www.bbc.com/news/technology-27711109`. Acessado em 12, 2015.

Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy";-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, New York, NY, USA. ACM.

Cha, M., Benevenuto, F., Haddadi, H., and Gummadi, K. (2012). The world of connections and information flow in twitter. In *IEEE Transactions on Systems, Man and Cybernetics - Part A*.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Cheang, H. S. and Pell, M. D. (2011). Recognizing sarcasm without language: A cross-linguistic study of english and cantonese. page 19.

Gibbs, R. W. and Colston, H. L. (2007). *Irony in language and thought: A cognitive science reader*. Psychology Press.

Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM Web Science Conference (WebSci)*.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Kreuz and Glucksberg (1989a). How to be sarcastic: the echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, pages 374–386.

Kreuz, R. J. and Caucci, G. M. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kreuz, R. J. and Glucksberg, S. (1989b). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118(4):374.

Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.

Li, G., Ghosh, A., and Veale, T. Constructing a corpus of figurative language for a tweet classification and retrieval task.

Liebrecht, C., Kunneman, F., and van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013*, page 29.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *Machine Learning*, pages 285–318.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Int'l Conference on World wide web (WWW)*, pages 851–860.

Singh, R. K. (2012). Humour, irony and satire in literature. pages 65–72.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6:1453–1484.

Valitutti, R. (2004). Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.