Analise da Confiabilidade da Informação Propagada em Mídias Sociais

Thiago Garcia Moreira, Juliana Valério, Jonice Oliveira,

Universidade Federal do Rio de Janeiro (UFRJ)

thiago.moreira@ppgi.ufrj.br, juvianna@dcc.ufrj.br, jonice@dcc.ufrj.br

Resumo. *As mídias sociais são usadas pra compartilhar informações e expressar opiniões sobre eventos. Mas nem todo conteúdo gerado é confiável. Este trabalho realiza um estudo das características que esse tipo de mensagem possui. Especificamente são analisadas postagens no Twitter, onde foram escolhidos sete rumores que se propagaram pelo Brasil entre julho de 2014 e janeiro de 2015. Esta pesquisa é base para um método para identificar inconsistência de informação nas mídias sociais.*

1. Introdução

Com a evolução das redes sociais on-line, três grandes mudanças ocorreram em relação ao uso da internet: i) a internet começa a substituir mídias tradicionais de fontes de notícias como jornais e televisão (KWAK, LEE, *et al.*, 2010); ii) a internet provê uma plataforma comum para as pessoas compartilharem informação e expressarem suas opiniões; iii) o aumento do número de dispositivos móveis permitiu que pessoas se conectem e usem seus aplicativos a qualquer momento, mudando o comportamento social (SRIVASTAVA, 2005). Além disso, o curto tempo de resposta e conexão de alta velocidade permite a disseminação rápida em larga escala de conteúdo.

A diferença entre a disseminação da informação nas mídias sociais e nos veículos tradicionais é que na primeira, o conteúdo é gerado pelos usuários. Devido ao anonimato provido pela internet e a alta velocidade de propagação da informação nessas mídias, um conteúdo não crédulo pode propagar rapidamente (DOERR, FOUZ e FRIEDRICH, 2012). Como milhões de postagens são geradas por hora, é difícil de monitorar qual conteúdo pode ser considerado verídico e qual não, e identificar as fontes de propagação dos rumores é ainda outro problema.

A propagação de rumores pode causar diversos prejuízos, como danos físicos e morais a pessoas, pânico em parte da população e atrapalhar situações de emergência (como resgates ou reconhecimento de eventos). Por exemplo, a propagação da falsa informação sobre o término da Bolsa Família gerou pânico em parte da população e ocasionou a lotação das agências da Caixa Econômica Federal. Temos ainda o caso onde a conta no *Twitter* do conselheiro presidencial da gestão de desastres da Indonésia foi invadida expondo falsos alertas de Tsunamis. Também o caso de uma mulher morta por espancamento após boatos, em mídias sociais, de sua participação em cultos satânicos. Portanto, identificar rumores nas mídias sociais é importante para evitar as consequências danosas que este tipo de conteúdo pode disseminar.

O objetivo deste trabalho é elencar um conjunto de fatores úteis para a analise da estrutura sintática das mensagens, suas semânticas e estrutura do perfil dos usuários, visando à identificação de rumores propagados em mídias sociais. Por analise sintática

entendemos como o estudo da forma gramatical que esses textos são escritos. A analise semântica caracteriza-se como o estudo do sentimento expressado por esses textos e pelos textos em resposta a eles. Por analise de perfil definimos o estudo das características comuns de usuários que postam rumores possuem, como idade, número de seguidores, data de criação do perfil, entre outros. Através deste conjunto de fatores, visa-se identificar padrões que possam colaborar para uma possível automação na identificação de inconsistência de informação. O foco deste trabalho é nas informações em português-brasileiro, disseminadas pelo *Twitter*.

2. Trabalhos Relacionados

Uma das maneiras de se identificar tais inconsistências é através da analise da estrutura da mensagem. Este estudo consiste em identificar padrões que possam caracterizar um conjunto de mensagens. Estes estão divididos em três grupos definidos a seguir:

- Mensagem: Analise sintática da mensagem disseminada.
- Usuário: Analise das características do perfil do usuário que disseminou a mensagem. Por analise de perfil definimos o estudo das características comuns de usuários que postam rumores possuem, como idade, número de seguidores, data de criação do perfil, entre outros.
- Resposta Social: Analise sintática e semântica do texto postado em resposta a mensagem original, como *retweets* e comentários.

O trabalho de Castillo *et al.* (2011) utiliza técnicas de aprendizado supervisionado de máquina para classificar um tópico no *Twitter* como notícia ou conversa pessoal e dentro da classe notícia classificá-la como crédula ou não crédula. A cla*ssificaç*ão é feita por humanos e um algoritmo de aprendizado extrai os padrões que *tweets* dessas classes possuem. O trabalho conclui que tópicos de notícias costumam possuir *links* (normalmente para a fonte da notícia) e uma árvore de propagação maior. Notícias crédulas são propagadas por autores que já possuem muitos *posts*, originam de um ou poucos usuários na rede e possuem grande quantidade de *retweets*. Por outro lado, as notícias não crédulas seriam propagadas por usuários que faz muito uso de *emoticons*, possuem poucos amigos e possuem menos tempo de atividade na rede.

Seguindo uma linha um pouco diferente Guoyong Cai *et al.* (2014) muda a premissa prévia de analisar o conteúdo em questão e passa a estudar a resposta social a esses conteúdos postados. É assumido em sua pesquisa que *retweets* e comentários correspondem a respostas sociais que podem descrever a credibilidade da informação inserida na rede. As principais diferenças deste trabalho aos que já foram desenvolvidos utilizando analise sintática podem ser resumidas da seguinte forma:

- Não existem estudos anteriores para rumores escritos em português-brasileiro, tornando-se um material importante para o cenário nacional. É de importância conhecer padrões que rumores possuem, pois podem ser úteis para automatizar uma previsão desse tipo de mensagem.
- Esta pesquisa estuda tanto a analise da estrutura da informação não crédula quanto da resposta social a essa informação. As pesquisas anteriores que utilizaram analise sintática, não incluíram este tipo de analise nas respostas das mensagens.

- Esta pesquisa realiza analise do perfil do usuário fonte, o primeiro a disseminar a mensagem no evento. Essa analise é importante, pois permite identificar padrões de características que possam possuir.

3. Método

3.1 - Identificação dos atributos a serem analisados

Como visto anteriormente, o estudo da estrutura consiste em identificar padrões de três grupos: Mensagem, usuário e resposta social. Atributos relacionados às características da mensagem podem ser dependentes ou independentes do *Twitter*. Independentes do *Twitter* existem em outras mídias além do *Twitter*, são eles: tamanho do texto, pontuação, fração de letras maiúsculas, uso de *smiles*, presença de URL e analise de sentimento. Dependentes do *Twitter* existem apenas no Twitter, são eles: Presença de menção, *hashtag* e se a mensagem é um *retweet*.

Atributos relacionados ao usuário consideram características dos usuários como: data de registro do perfil, número de seguidores, número de amigos, e quantidade de mensagem que ele já realizou até o momento. Os atributos escolhidos para o estudo da estrutura do texto são os mesmos estudados por Castillo *et al.* (2011), com a adição da distinção entre *smiles* postivos e negativos a fim de idenfiticar uma possível polaridade que possa caracterizar um padrão. Nos baseamos nos atributos de Castillo *et al.* (2011) porque nele foi realizada uma analise sintática completa atingindo resultados satisfatórios na classificação automática da credibilidade da informação. A Tabela 1 ilustra os atributos analisados e o que representam.

Tabela 1: Atributos escolhidos para analise e sua descrição.

Grupo	Atributo	Descrição
Usuário	*Contagem de mensagens*	Quantidade de mensagens (*tweets)* disseminadas pelo usuário até a mensagem analisada
	Seguidores	Quantidade de seguidores
	Amigos	Quantidade de amigos, usuários seguidos.
	Idade	Tempo de existência do perfil medido em anos
Mensagem	*(*)Presença de URL*	Frequência do uso de URLs nas mensagens. Quantidade de URL
	(*)Menção	Frequência de menções na mensagem, quando há uma referência a outro usuário usando o prefixo "@".
	(*)!	Frequência de exclamação na mensagem. Quantidade de "!" pelo tamanho do texto.
	(*)?	Frequência de interrogação na mensagem. Quantidade de "?" pelo tamanho do texto.
	(*)Smiles positivos	Quantidade de *smiles* postivos na mensagem.
	(*)Smiles negativos	Quantidade de *smiles* negativos na mensagem
	(*)Fração caixa alta	Frequência de caracteres em caixa alta na mensagem. Quantidade de caracteres em caixa alta pelo tamanho do texto.
	(*)Hashtag	Quantidade de *hastags* na mensagem
	(*)RT	Se a mensagem é um *Retweet*
Resposta social	*Frequência Mensagem Positiva*	Mensagem expressando sentimento positivo
	Frequência Mensagem Negativa	Mensagem expressando sentimento negativo

Para cada atributo, marcado com *(*),* descrito na Tabela 1, são calculados a frequência de sua aparição em relação à quantidade de mensagens em cada base. Essa métrica (Equação 1) é definida com o intuito de identificar um padrão. Onde "F" representa a frequência do atributo no *corpus*, "f" a frequência do atributo em um *tweet* e "q" a quantidade total de *tweets*.

$$F = (\sum_{0}^{q} f)/q$$

Equação 1: Frequência

Para a analise da Resposta Social, analisaram-se os *retweets* (RT) de mensagens não crédulas. Nesta analise, aplicam-se as mesmas métricas apresentadas na Tabela 1, além da analise de sentimento sobre os RT. Este trabalho adiciona o sentimento neutro, não utilizado por nenhum outro trabalho citado anteriormente para analise de credibilidade da informação. Decidimos incluir a analise do sentimento neutro para evitar classificar frases que não expressam sentimento em positiva ou negativa. A analise de sentimento é constituída por duas etapas, são elas:

- Pré-processamento: É removido do texto as pontuações, palavras de parada, prefixos de URI e os caracteres específicos do *Twitter* (@ , # e RT). Foi utilizada a biblioteca NLTK em português.
- Algoritmo de Classificação: Foi utilizado o classificador *Naive Bayes* da biblioteca NLTK treinado com um conjunto de palavras positivas e outro de palavras negativas. Como resultado o classificador indica a probabilidade do texto possuir sentimento que foi classificado, positivo, negativo ou neutro.

3.2. Coleta de Dados

Foram coletados dados sobre sete falsas notícias que propagaram pelo *Twitter* entre os meses de julho de 2014 e janeiro de 2015: i) a morte de Jô Soares; ii) fusão das lojas Marisa e Renner; iii) Vírus ebola no Brasil; iv) Aposentadoria de Silvio Santos; v) Saída do jogador Daniel Alves do Barcelona; vi) Motivo do afastamento do jogador Maicon da seleção vii) Morte do cantor Jacaré. Para a coleta foi utilizado a API 1.1 do *Twitter*.

Foram coletadas no total 39.209 mensagens realizadas por 26.570 usuários. Todos os dados foram armazenados no MongoDB, por se tratar de um banco NoSQL, facilitando a manipulação dos arquivos json retornados pela API do *Twitter*.

4. Analise dos Resultados

Aplicando as métricas definidas na Tabela 1 aos dados coletados, obtiveram-se os seguintes resultados ilustrados na Tabela 4, que apresenta o valor da Equação 1 aplicada em cada um dos rumores.

Na Tabela 2 observa-se que a presença de URL foi alta em todos os rumores exceto no sexto, provavelmente por este ter sido gerado em comentários de uma mídia social (Facebook) e não em uma notícia, onde essa característica é mais comum (CASTILLO, MENDOZA e POBBLETE, 2011). A presença de menções também foi alta em todos os rumores, exceto no primeiro e terceiro, provavelmente por estes possuírem o menor apelo social. A frequência de interrogação e exclamação foi alta em todos os rumores, com seus picos no primeiro, sexto e quarto, onde estão relacionados a pessoas famosas. O uso de *smiles* foi baixo em todos os rumores. A frequência de caracteres em caixa alta não teve grande variação, tendo seu auge alcançado no rumor 3, relacionado a duas grandes empresas. O uso de *hastag* também se manteve constante e baixo. Por último, a presença de *retweets* foi alta em todos eles, indicando uma preferência pela forma com que usuários disseminam informação não crédula.

Ainda na Tabela 2 observa-se que na parte dos usuários participantes, apesar da baixa idade do perfil, a quantidade de mensagens que já disseminaram é alta. Também são usuários que possuem muitos seguidores e poucos amigos (seguidos). Já os usuários fonte, possuem idade de perfil acima da média, baixa razão entre amigos e seguidores, e um número elevado de mensagens acumuladas (contagem de mensagem).

Tabela 2: Equação 1 aplicada individualmente a todos os rumores.

Rumores								
Atributos		**1-Jo**	**2-Daniel**	**3-Marisa**	**4-Jacaré**	**5-Silvio**	**6-Maicon**	**7-Ebola**
Usuário fonte	Contagem de mensagem	3.254	28.441	3.305	17.286	8.819	62.017	164.160
	Seguidores	62	441	1.428	6.664	1.939	52.012	995
	Amigos(Seguidos)	53	1.480	1.400	0	1.931	383	659
	Seguidos/seguidores	0,83	3,35	0,9	0	0,9	0,011	0,662
	Idade	4	5	3	5	4	2	4
Usuário participante	Contagem de mensagem	37.249,0	64.611,0	40.693,0	31.502,0	44.180,0	30.882,0	38.488,0
	Seguidores	14.405,0	3.603,0	10.228,0	1.496,0	11.134,0	2.577,0	8.715,0
	Amigos(Seguidos)	1.036,0	1.172,0	1.397,0	637,0	1.251,0	803,00	1.182,0
	Seguidos/seguidores	0,07	0,33	0,13	0,43	0,11	0,311	0,13
	Idade	2,70	2,48	0,80	4,22	2,83	3,95	2,88
Texto	Presença de URL	0,73	0,95	0,44	0,96	1,00	0,26	0,61
	Menção	0,52	0,25	0,12	0,50	0,44	0,73	0,61
	!	0,12	0,05	0,09	0,26	0,06	0,14	0,08
	?	0,15	0,06	0,01	0,36	0,16	0,45	0,08
	Smile Pos	0,0075	0,00	0,0023	0,0219	0,0032	0,0043	0,0023
	Smiles Neg	0,0054	0,00	0,0390	0,0305	0,0040	0,0038	0,0018
	Frac Caixa alta	0,06	0,03	0,21	0,09	0,05	0,06	0,04
	Hashtag	0,13	0,16	0,24	0,16	0,08	0,09	0,16
	RT	0,33	0,13	2,35	0,42	0,34	0,57	0,51

Analisando apenas a resposta social a essas postagens, como os dados vieram do *Twitter*, a forma de verificar essa resposta é analisando os RT. A Tabela 3 apresenta a média juntando a resposta social de todos os RT.

Tabela 3: Resultados da analise da resposta social

RT	F
!	0,1707
?	0,1613
Smile Pos	0,0023
Smiles Neg	1,0102
Frac Caps Lock	0,0623

Observa-se que o uso de *smiles* positivos é muito baixo (0,2%), enquanto que o uso de *smiles* negativos é elevado. O uso de exclamações e interrogações é a cima do normal, assim como a frequência de letras maiúsculas na frase.

A analise de sentimento feita está ilustrada na Tabela 4. A classificação foi feita de acordo com três conjuntos, sentimento positivo, sentimento negativo e sentimento neutro. Os trabalhos correlacionados que fazem esse tipo de analise em credibilidade da informação ainda não estudaram a neutralidade.

Tabela 4: Analise de Sentimento

Total					
		Todos *Tweets*	%	Apenas RT	%
Sentimento	POSITIVOS	7.742	19,75	4.029	21,19
	NEUTROS	18.523	47,24	9.630	50,64
	NEGATIVOS	12.944	33,01	5.359	28,18
	TOTAL	39.209	100	19.018	100

De acordo com a analise de sentimento realizada é possível concluir que pessoas não demonstram sentimento positivo com postagens de rumores: 50% dos RT dessas postagens apresentaram um sentimento neutro e 28% apresentaram um sentimento negativo. Esta analise está de acordo com a analise sintática da resposta social, onde foi observado o uso de *smiles* negativos em mais de 100% dos *tweets*. Os *smiles* não foram levados em consideração na analise de sentimento.

5. Conclusão e Trabalhos Futuros

Este trabalho realizou um estudo utilizando a estrutura de rumores no cenário brasileiro. Constatou-se que para o cenário brasileiro o uso de *smiles*, não pode ser utilizado para prever incredibilidade da informação. Porém na analise de resposta social (especificamente os *retweets*), observa-se um alto índice na frequência de *emoticons* negativos (maior que 100%) e um baixo índice no uso dos *emoticons* positivos (6%).

Este trabalho observou o comportamento dos usuários brasileiro propagadores de rumores. No cenário brasileiro, o uso de URLs atrelados com menções é comumente utilizado (provavelmente na intenção de propagar a notícia a outras pessoas.). Como trabalhos futuros, pode-se analisar, além dos RT, as conversas que ocorrem no *Twitter* sobre uma mensagem não crédula, pois é outra forma de resposta social sobre um *tweet*. Também seria interessante uma analise independente do idioma utilizado. Pretende-se enriquecer a analise de conteúdo e inserir a analise de propagação da informação. Esta pesquisa pode ser aplicada a outros rumores que se propagarem pelo cenário brasileiro a fim de obter resultados mais precisos sobre os padrões das postagens nesse cenário.

Referências

CASTILLO, C.; MENDOZA, M.; POBBLETE, B. Information Credibility on Twitter. **WWW '11 Proceedings of the 20th international conference on World wide web**, 2011.

DOERR, B.; FOUZ, M.; FRIEDRICH, T. Why rumors spread so quickly in social networks. **Communications of the ACM**, 2012.

GUOYONG, C.; HAO, W.; RUI, L. Rumors Detection in Chinese via Crowd Responses. **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)**, 2014.

KWAK, H. et al. What is Twitter, a social network or a news media? **WWW'10**, 2010.

SRIVASTAVA, L. Mobile phones and the evolution of social behavior. **Behaviour & information technology**, 2005. 111-119.