Caracterização da Propagação de Rumores no Twitter utilizando Redes Textuais Temporais

Fabíola S. F. Pereira[1]

[1]Faculdade de Computação
Universidade Federal de Uberlândia (UFU)
Uberlândia – MG, Brasil

fabiola.pereira@ufu.br

Abstract. *Social media platforms are the main means of disseminating information and news, being valuable in many applications. At the same time, they are platforms that enhance the spread of rumors and fake news. Three fundamental elements for understanding the information exchanged between people on social networks are the individuals or actors, the information they exchange and the time when such information is exchanged. This article presents an analysis of the spread of rumors using structures of temporal textual networks. The results indicate that such structures are ideal for analyzing the problem of spreading rumors on Twitter.*

Resumo. *Plataformas de mídias sociais são o principal meio de difusão de informação e notícias, sendo valiosas em muitas aplicações. Ao mesmo tempo, são plataformas que potencializam a difusão de rumores e notícias falsas. Três elementos fundamentais para entender a informação trocada entre pessoas em redes sociais são os indivíduos ou atores, a informação que eles trocam e o tempo quando tais informações são trocadas. Neste artigo é apresentada uma análise da propagação de rumores utilizando estruturas de redes textuais temporais. Os resultados indicam que tais estruturas são ideais para análise do problema de propagação de rumores no Twitter.*

1. Introdução

Rumores são informações não-verificadas no momento da publicação [Zubiaga et al. 2016]. Notícias de última hora são clássicos cenários de geração de rumores, já que elas levam a situações de publicação em ritmo acelerado em mídias sociais, produzindo muitas atualizações que, na maioria das vezes, não são verificadas.

A tarefa de detecção de rumor [Zubiaga et al. 2018] consiste em distinguir pedaços de informação não-verificados (rumores) de todo o restante da informação em circulação (não-rumores). Tal tarefa é especialmente importante, pois assinalar uma publicação como não-verificada ajuda a evitar a propagação de informações que podem vir a se tornar falsas (*fake news*).

Muitas abordagens têm surgido com o objetivo de resolver a tarefa de detecção de rumores em mídias sociais. Características textuais relacionadas ao estilo de escrita da publicação [Zubiaga et al. 2018, Kotteti et al. 2020], características sociais do usuário que publica a informação [Kochkina et al. 2018] e observação da sequência temporal de publicação [Zubiaga et al. 2018], são exemplos das informações que têm sido exploradas

– muitas vezes separadamente – na tarefa de detecção de rumores. Nenhum trabalho, entretanto, explorou simultaneamente (i) a sequência temporal das publicações, (ii) o percurso das mensagens na rede e (iii) o conteúdo textual das mensagens.

Redes textuais temporais são modelos de representação que codificam os componentes chave das redes de informação (topologia, texto e tempo) em uma única rede bipartite, de maneira que seja possível representar diferentes formas de comunicação [Vega and Magnani 2018]. Um modelo de rede textual temporal pode ser analisado direta ou indiretamente para realizar diferentes tarefas de mineração de dados.

Neste trabalho o objetivo é utilizar a estrutura de redes textuais temporais para modelar o problema de propagação de rumores no Twitter. A topologia temporal de tais redes representa a cadeia de influência que uma mensagem sofre ou exerce ao longo do tempo (modelada neste trabalho por meio de menções entre usuários). Assim, partindo-se do princípio de que rumores surgem essencialmente de publicações em ritmo acelerado em mídias sociais, torna-se interessante observar o comportamento da propagação de uma mensagem publicada na rede sob tais condições.

Uma vez modelado o problema, é também objetivo caracterizar a propagação das mensagens utilizando métricas extraídas da rede textual temporal. Tais métricas permitem entender como é a persistência, disseminação e homogeneidade dos caminhos iniciados por mensagens na rede. Por fim, trata-se de um trabalho em andamento cujo objetivo final é incorporar em algoritmos de detecção de rumores, padrões de propagação de rumores em redes sociais. A análise exploratória reportada neste artigo auxilia na coleta de indícios para responder à seguinte questão: *rumores propagam de maneira diferente de mensagens que não são rumores em redes sociais?*

2. Trabalhos Correlatos

Não são muitos os trabalhos dedicados à tarefa de detecção de rumor [Zhao et al. 2015, Zubiaga et al. 2016, Kochkina et al. 2018, Kotteti et al. 2020, Yang et al. 2020]. Uma das primeiras abordagens foi introduzida em [Zhao et al. 2015], baseada em regras para identificar o ceticismo da mensagem e, portanto, determinar se a informação é um rumor. As limitações dessa abordagem consistem em ter que esperar por respostas ao invés de uma identificação em tempo real. Além do mais, há a falta de generalização já que as regras são definidas manualmente. [Zubiaga et al. 2016] propôs uma abordagem capaz de aprender a dinâmica da informação durante notícias de última hora, classificando um texto como rumor ou não-rumor considerando o contexto aprendido à medida que o evento vai acontecendo. Ou seja, considerando a sequência de *tweets* como observada na linha do tempo do Twitter. Os trabalhos mais recentes baseiam-se na abordagem de problema proposta em [Zubiaga et al. 2016], sendo algoritmos essencialmente focados em aprendizado profundo [Kochkina et al. 2018, Kotteti et al. 2020, Yang et al. 2020].

Não há consenso na literatura em relação à definição de rumores. Muitas vezes notícias falsas, desinformação, rumores, *spam* e lenda urbana são termos tratados como sinônimos [Wu et al. 2019]. Nesse sentido, há uma vasta literatura dedicada especialmente à caracterização da propagação de notícias falsas nas redes sociais [Pierri and Ceri 2019, Vosoughi et al. 2018]. Neste trabalho, entretanto, rumor é definido como um pedaço de informação não-verificada, podendo ser verdadeira ou falsa.

Dentre os trabalhos relacionados à utilização de redes textuais temporais, em

[Vega and Magnani 2018] é feita uma análise sobre dados do Twitter, acerca da evolução dos tópicos publicados na rede relacionados ao tema IoT. [Vega and Magnani 2019] mostram a aplicabilidade do modelo em um estudo sobre uma base de dados também do Twitter contendo publicações de políticos da Suécia durante o período de 1 mês. Apenas observações sobre as características dos caminhos encontrados foram reportadas.

3. Propagação de Mensagens em Redes Textuais Temporais

De acordo com [Vega and Magnani 2018], uma rede textual temporal é um grafo $G = (A, M, E)$ direcionado bipartite representando uma rede de comunicação, onde A é o conjunto de nós do tipo atores, M é o conjunto de nós do tipo mensagens e E é o conjunto de arestas anotadas por um instante de tempo t. A direção das arestas indica o fluxo de comunicação: $e_1 = (a_i, m_k, t_x) \in E$ indica que o ator a_i produziu o texto m_k (a_i é produtor) no tempo t_x e $e_2 = (m_k, a_j, t_y) \in E$ indica que a_j consumiu a mensagem m_k no tempo t_y (a_j é consumidor), para $t_x \leq t_y$. Diz-se que e_1 é incidente em e_2. A Figura 1 ilustra uma rede textual temporal.

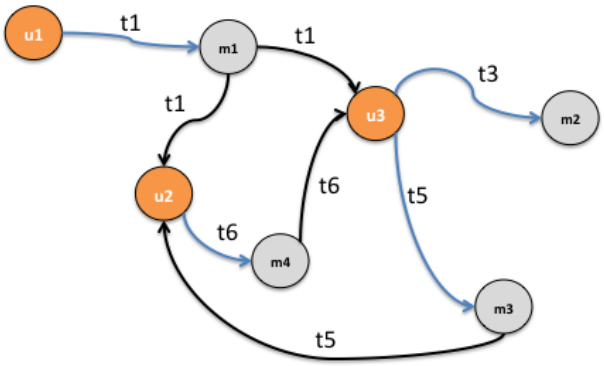

Figura 1. Modelo de uma rede textual temporal. Os nós são do tipo atores (laranja) e mensagens (cinza), ligados por arestas temporais, indicando os instantes em que mensagens são produzidas (azul) e consumidas (preto) por atores.

Um caminho temporal p é uma sequência de arestas $e_1, e_2, ..e_n$, onde e_i é incidente em e_{i+1} para todo $1 \leq i \leq n-1$, tal que nenhum nó é repetido.

No contexto de propagação de rumores, o interesse está na análise dos caminhos temporais entre duas mensagens (informação). Ou seja, a partir de uma mensagem deseja-se caracterizar a propagação e potencial de transformação de tal mensagem na rede. Para tanto, as seguintes métricas, originalmente propostas em [Vega and Magnani 2019], são adaptadas neste trabalho para o contexto de modelagem de rumores. Considere $p(m, s)$ um caminho temporal tal que $m, s \in M$.

- **Disseminação** $D(m)$**.** Seja o tamanho topológico de p o número de atores em p. $D(m)$ é a média dos tamanhos topológicos de todos os caminhos com origem em m, para todo $s \in M$.
- **Persistência** $P(m)$**.** Seja o tamanho temporal de p a diferença de tempo entre a última aresta produtora e o tempo da primeira aresta consumidora. $P(m)$ é a média dos tamanhos temporais de todos os caminhos com origem em m, para todo $s \in M$.

- **Entropia** $H(m)$**.** Dado um caminho p, tem-se $\rho_i(p) = \frac{\sum_{m \in p} c_i(m)}{M_p}$, onde M_p é o número de mensagens em p e $c_i(m)$ uma função que mapeia a mensagem m em uma classe c_i para $1 \leq i \leq n$, onde n é o número de classes. A entropia textual de um caminho p é definida como:

$$H(p) = -\sum_{i=1}^{n} \rho_i(p) ln \rho_i(p) \tag{1}$$

  $H(m)$ é a média das entropias textuais de todos os caminhos com origem em m para todo $s \in M$.

Disseminação refere-se à quantidade de usuários que consomem e produzem mensagens em um caminho temporal. A persistência mede o tempo em que mensagens perduram na rede, impactando usuários a produzirem novas mensagens. E a entropia descreve o quão homogêneas são as mensagens produzidas e consumidas em sequência temporal em relação à classe a qual pertencem.

4. Modelando Rumores no Twitter como uma Rede Textual Temporal

Duas bases de dados foram analisadas, obtidas de [Zubiaga et al. 2016]. Trata-se de dados coletados do Twitter, com *tweets* referentes a notícias de última hora, anotados como rumor ou não-rumor. O processo de anotação emulou o cenário em que um usuário está acompanhando informações associadas com notícias de última hora (*breaking news*). Apenas observando a linha do tempo de tais notícias através do Twitter, o usuário deve ser capaz de decidir se um novo *tweet* é um rumor ou não.

A primeira base de dados refere-se ao evento do "ataque à Charlie Hebdo", ataque terrorista ao jornal Charlie Hebdo, ocorrido em Paris, 2015, contendo 2079 *tweets* divididos entre 22% de rumores e 78% de não-rumores. A segunda base de dados refere-se ao evento do "cerco de Sydney", ataque terrorista ocorrido em Sydney, Austrália, 2014, com 1221 *tweets*, sendo 42.8% rumores e 57.2% não-rumores. A escolha de tais bases busca garantir diversidade quanto à distribuição das classes (rumor ou não-rumor). Enquanto a primeira é desbalanceada, a segunda possui menor desproporcionalidade entre as classes.

A rede modela a propagação dos *tweets* entre os usuários com base nas menções. Menções são *tweets* que incluem outro usuário do Twitter em seu conteúdo através do símbolo @. Existem dois tipos de nós: atores (usuários) e mensagens (*tweets*). Uma aresta de um ator a para um mensagem m em um instante t é uma aresta *produtora*, indicando que a publicou m em t. Enquanto uma aresta de uma mensagem m para um ator b em um instante t é chamada de *consumidora*, indicando que m tem uma menção ao ator b e foi publicada em t. Trata-se, portanto, de uma rede bipartida. Note que nesta modelagem parte-se do princípio que o ator mencionado está recebendo a mensagem e que ele a consome no mesmo instante em que foi publicada, o que pode não necessariamente ser verdade - por exemplo, usuários não estão conectados o tempo todo no Twitter. Assim, é uma modelagem que considera o cenário mais rápido possível em termos de tempo de consumo de mensagens.

A Figura 2 ilustra parte da rede `charlie-hebdo-net` em três dias diferentes. O primeiro dia corresponde ao dia do ataque terrorista. À medida que o tempo passa, mais rumores são publicados. É possível identificar visualmente também os usuários que

mais publicam - nós em cor laranja com alto valor de grau de saída (número de arestas com origem no nó).

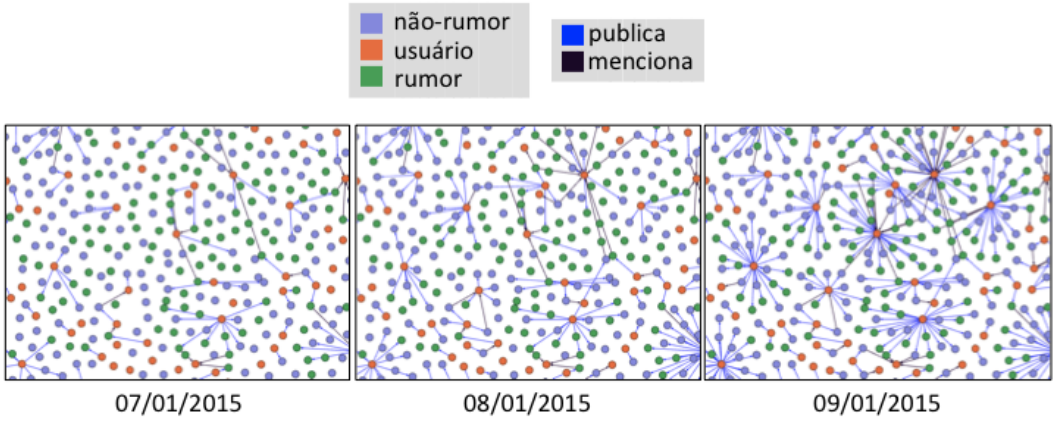

Figura 2. Evolução da rede textual temporal `charlie-hebdo-net`. Os nós não estão representados de maneira dinâmica para melhor comparação visual entre instantes de tempo.

5. Análise Exploratória

A tabela 1 sumariza as características das redes textuais temporais analisadas. A rede `charlie-hebdo-net` é maior e mais densa. A rede `sydney-net` contém os *tweets* relativos a um *timespan* menor que 1 dia. Em ambas redes, arestas do tipo "menciona" (consumidoras) correspondem a menos de 15% do total de arestas.

Tabela 1. Estatísticas das redes textuais temporais analisadas.

		`charlie-hebdo-net`	`sydney-net`
arestas	publica	2079	1221
	menciona	339	171
	total	2418	1392
nós	usuários	1191	633
	rumores	458	522
	não-rumores	1621	699
	total	3270	1854
timespan		17/01/2015 a 19/01/2015	14/12/2014 21hs a 15/12/2014 13hs

Com base nas métricas propostas, para cada mensagem m de uma rede foram calculadas a disseminação $D(m)$, persistência $P(m)$ e entropia $H(m)$. A tabela 2 sumariza os resultados, contendo o valor médio entre os nós do tipo mensagem para cada métrica. Os resultados indicam que há uma diferença de comportamento entre mensagens do tipo rumor e mensagens do tipo não-rumor.

Tanto na rede `charlie-hebdo-net` quanto na rede `sydney-net` o padrão se manteve: rumores têm a característica de gerar caminhos com menor persistência e disseminação. Além do mais, rumores geram caminhos com alta entropia, ou seja, caminhos heterogêneos. Por outro lado, notícias de última hora que não são rumores têm a

característica de maior disseminação e maior persistêncai na rede, além de darem origem a caminhos mais homogêneos.

Tabela 2. Características das mensagens das redes analisadas. P medida em horas.

	charlie-hebdo-net			sydney-net		
	D	P	H	D	P	H
rumor	1.15	1.23	0.69	1.01	0.33	0.78
não-rumor	1.87	4.45	0.15	1.34	0.87	0.23

6. Considerações Finais

Redes textuais temporais têm-se mostrado boas estruturas para modelagem da propagação de rumores. Neste trabalho foi proposta uma modelagem inédita de notícias de última hora coletadas do Twitter e classificadas como rumores ou não. Com a modelagem proposta, foi possível explorar três características das mensagens trocadas na rede: persistência, disseminação e entropia. São características que descrevem informações topológicas, temporais e de conteúdo trafegado em uma rede social.

Com as métricas apresentadas neste trabalho, foi possível realizar uma análise exploratória do comportamento da rede social em relação a notícias de última hora, permitindo quantificar, de maneira inédita, o impacto temporal de rumores no contexto analisado.

Os experimentos também mostraram indícios de que rumores propagam de maneira diferente de mensagens que não são rumores. É um trabalho em andamento a utilização das métricas aqui propostas para serem incorporadas em algoritmos para detecção de rumores.

Referências

Kochkina, E., Liakata, M., and Zubiaga, A. (2018). All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413. Association for Computational Linguistics.

Kotteti, C. M. M., Dong, X., and Qian, L. (2020). Ensemble deep learning on time-series representation of tweets for rumor detection in social media. *CoRR*, abs/2004.12500.

Pierri, F. and Ceri, S. (2019). False news on social media: A data-driven survey. *SIGMOD Rec.*, 48(2):18–27.

Vega, D. and Magnani, M. (2018). Foundations of temporal text networks. *Applied Network Science*, 3(25).

Vega, D. and Magnani, M. (2019). *Metrics for Temporal Text Networks*, pages 147–160. Springer International Publishing, Cham.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 6380(359).

Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Yang, X., Lyu, Y., Tian, T., Liu, Y., Liu, Y., and Zhang, X. (2020). Rumor detection on social media with graph structured adversarial learning. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1417–1423. International Joint Conferences on Artificial Intelligence Organization.

Zhao, Z., Resnick, P., and Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1395–1405. International World Wide Web Conferences Steering Committee.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

Zubiaga, A., Liakata, M., and Procter, R. (2016). Learning reporting dynamics during breaking news for rumour detection in social media. *CoRR*, abs/1610.07363.