Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Early Detection of Rumor Veracity in Social Media

Anh Dang
Dalhousie University
Halifax, NS
Canada B3H 4R2
anh@cs.dal.ca

Abidalrahman Moh'd
Eastern Illinois University
Charleston, Illinois
USA 61920
mabidalrahman@eiu.edu

Aminul Islam
University of Louisiana at Lafayette
Lafayette, LA
USA 70503
aminul@louisiana.edu

Evangelos Milios
Dalhousie University
Halifax, NS
Canada B3H 4R2
eem@cs.dal.ca

Abstract

Rumor spread has become a significant issue in online social networks (OSNs). To mitigate and limit the spread of rumors and its detrimental effects, analyzing, detecting and better understanding rumor dynamics is required. One of the critical steps of studying rumor spread is to identify the level of the rumor truthfulness in its early stage. Understanding and identifying the level of rumor truthfulness helps prevent its viral spread and minimizes the damage a rumor may cause. In this research, we aim to debunk rumors by analyzing, visualizing, and classifying the level of rumor truthfulness from a large number of users that actively engage in rumor spread. First, we create a dataset of rumors that belong to one of five categories: "False", "Mostly False", "True", "Mostly True", and "Half True". This dataset provides intrinsic characteristics of a rumor: topics, user's sentiment, network structural and content features. Second, we analyze and visualize the characteristics of each rumor category to better understand its features. Third, using theories from social science and psychology, we build a feature set to classify those rumors and identify their truthfulness. The evaluation results on our new dataset show that the approach could effectively detect the truth of rumors as early as seven days. The proposed approach could be used as a valuable tool for existing fact-checking websites, such as Snopes.com or Politifact.com, to detect the veracity of rumors in its early stage automatically and educate OSN users to have a well-informed decision-making process.

1. Introduction

Online rumors are truth-unverifiable statements or news that are spread and discussed in Online Social Networks (OSNs). They commonly appear and are propagated in uncertain situations [1]. Recently, social media has been used as a means to transmit information, such as breaking news, sport events, and political statements [2]. Although social media provides a reliable way to spread information to a large population in a short time, it also has a critical drawback. For example, a lot of information in social media could be rumors that are spread maliciously. Rumors that are "False", "Mostly False", or "Half True" could cause a tremendously adverse effect on people's lives. This raises the questions of how to identify, validate, and debunk the truthfulness of rumors. Researchers have tried to: detect rumors [3, 4], detect the original sources of rumors [5], identify who spread rumors [6], and rumor stance classification [7]. However, little work has been done to debunk and validate the level of the rumor truthfulness.

Studying rumor spread is an inherently interdisciplinary field. Social scientists have studied the intrinsic characteristics of rumor spread since the 1940s [8] and proposed theories on how rumors are propagated. For example, "The Basic Law of Rumor" [9] stated that the popularity of a rumor depends on the importance of its topic and the verifiability of its truthfulness. Recently, with advances in visualization and machine learning, it has become possible to apply knowledge from social science and psychology to better understand rumor spread in OSNs. Researchers have focused on the rumor veracity task [10, 11], i.e., given a rumor in social media and its related posts, to determine the veracity of this rumor (as "true", "false", or "unverified"). Most of the existing approaches tackle the problem from a machine learning point of view (e.g., trying various features and deciding what features produce the best result). These approaches may not be able to capture the changing characteristics of rumor spread [11]. In this paper, we explore if OSN user decision-making process in rumor spread could help to detect the level of rumor truthfulness in its early stage. Specifically, the approach analyzes what contributing factors play an important

URI: https://hdl.handle.net/10125/59674
ISBN: 978-0-9981331-2-6

role for a user to spread or debunk rumors. This not only provides more credible results but also explains whether social science theories could be applied to social media data.

Researchers have confirmed that false rumors, hoaxes, or fake news (another form of false rumors) are more prone to be spread further [12]. Those rumors have a tremendous effect on an individual's reputation or societies. For example, more than 50% of the voter population had seen fake news in the US 2016 election, and 50% of them believed in fake news stories [13]. Another example is rumors could play a detrimental effect on the stock markets [14]. Detecting rumor veracity in its early stage in OSNs is an essential step for end users to make a better-informed decision-making process. Recently, Google has teamed up with Snopes.com [15] and Politifact.com [16] to validate and debunk rumor stories in OSNs. Current approaches (e.g., Snopes.com and Politifact.com) use human knowledge to manually label if a rumor is "False", "Mostly False", "True", "Mostly True", and "Half True". As thousands of rumors are spread in OSNs in a very short time, manually labeling all those rumors is time-consuming and unrealistic. Recently, crowdsourcing solutions have been proposed to improve the results of machine learning tasks, such as machine translation [17] and sentiment analysis [18]. Some public crowdsourcing websites, such as Amazon Mechanical Turk, provide a mechanism to use human knowledge and insights to assign labels for some pre-defined tasks. However, each task has to be defined precisely, and the approach is not suitable for rumor labeling as rumors constantly change in real-time. To address this limitation, with a large number of users and their active participation, OSNs could effectively be a useful source of human input to debunk rumors [6]. Another limitation is that existing rumor veracity classification research only distinguishes if a rumor is "True", "False", and "Unverified". We propose a newly-created rumor dataset with finer-grained truth levels (according to Snopes.com and Politifact.com) and use this dataset to study how early we could effectively identify the truth of rumors.

This paper makes the following contributions:

- We introduce and analyze a dataset of 88 rumors from Reddit. Each rumor is identified as "False", "Mostly False", "True", "Mostly True", or "Half True". This dataset contains intrinsic characteristics of a rumor, such as topics, user's sentiment, network structural and content features.
- We demonstrate that using established social science and psychology theories helps to select better feature sets for the rumor veracity detection task and provides a better understanding and detection of rumor veracity by integrating these theories with visualizations and machine learning techniques.
- We evaluate how early we could effectively identify rumor truth values and provide insights into breaking news and long-standing rumors. Our experimental results show that we could effectively detect the truth of rumors as early as seven days.
- The proposed approach could bridge the gap between social science theories and experimental research of rumor spread in OSNs by validating and confirming if social science theories about user decision-making processes could be applied for social media data.

2. Related Work

2.1. Mining Rumor Data

One of the first publicly available rumor datasets is provided by [19]. This dataset includes 10,000 tweets involving five different rumors. Each tweet is annotated as "related" or "unrelated" to a rumor. A dataset of 100 million tweets involving 72 rumors (41 true and 31 false) was constructed by [20] and a machine learning approach was applied to it to classify whether those rumors are true or false. The PHEME dataset includes 1,972 rumorous and 3,830 non-rumorous tweets about five breaking news stories [21]. The dataset provided by [11] is a collection of tweets for 61 rumors and 51 non-rumors, used to study how various feature sets affect the accuracy of rumor detection over time. The SemEval 2017 Task 8 [10] provides a dataset that includes tweets and an annotation label for each tweet, "support", "deny", "query", or "comment". Eight teams participated and submitted the results for this task. The winning system classified the stance of each tweet using features and labels of the previous tweets. As most of the existing datasets only focus on "false" and "true" rumors, we aim to provide a rumor dataset that could be used to identify the truthfulness of rumors in one of the five categories: "False", "Mostly False", "True", "Mostly True", "Half True". These fine-grained truth levels are used to reflect the nature of rumor spread in OSNs.

2.2. Rumor Analysis in OSNs

One of the first analyses about rumor spread was in 1944 by [8]. The authors studied how rumors were spread in a particular neighborhood community. Due to the limitation of rumor data and the intrinsic

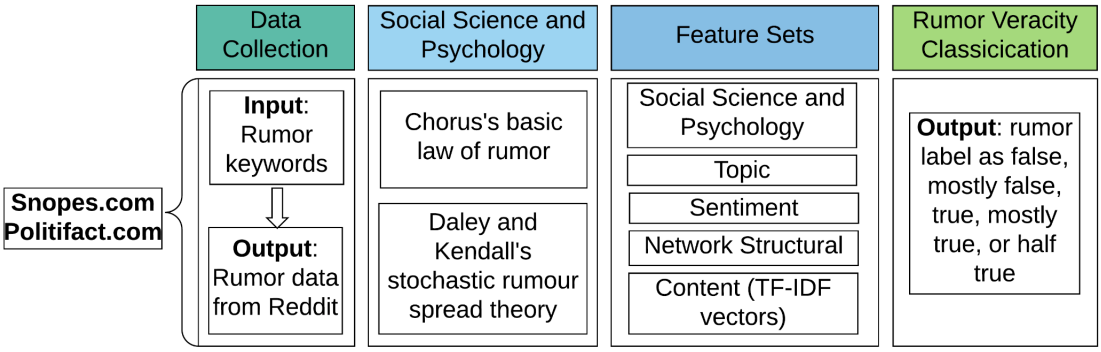


Figure 1: Rumor veracity classification framework.

long-lasting nature of rumors, rumor analysis was mostly theoretical research and experienced a long hiatus until the popularity of OSNs in the 2000s. In most OSNs, data is available, disseminated and stored permanently, so researchers have access to data to more efficiently study rumors and verify their theories. A classification approach to identify if a tweet is a rumor on Twitter was adopted by [19]. Each tweet was manually assigned as either being related to rumors or not. Relations between claims associated with rumors and analyzed contradictory claims inside a rumor were interpreted by [22]. The credibility perceptions of rumors were studied by classifying if a tweet is related to a rumor into three classes: "certain", "somewhat certain", and "uncertain" in [23]. An autonomous message-classifier that filters relevant and trustworthy tweets was proposed in [20]. How different feature sets could affect the performance of the rumor veracity task over time was studied in [11]. Our work is different from those approaches as our starting point consists of established theories from social science and psychology. Using those concepts, we propose a new rumor dataset that better reflects various truth levels of a rumor. For the classification task, we use various feature sets that are derived based on the notion of how rumors are transmitted in OSNs.

3. Methodology - Rumor Veracity Classification Task

The definition of rumor veracity classification was first proposed by [24]. However, the authors only consider three-class labels: "False", "True", and "Unverified". We extend that definition to our collected dataset with five class labels.

Let a rumor, RU_i, have a list of submissions in ascending time order SU_1, SU_2, ... SU_m and a list of topics T_1, T_2, ... T_n that are extracted from the submission titles in ascending time order. Each submission SU_j has a list of user comments C_1, C_2, ... C_k. Each user comment has a sentiment analysis score. If a comment C_x of user U_o replies to comment C_y of user U_p, there will be a connection from user U_o to user U_p in the user interaction graph. The task is to determine whether a rumor in OSNs could be categorized into one of the five categories: "False", "Mostly False", "True", "Mostly True", and "Half True" using the above submissions, user interaction graph, and other metadata. The architecture of the proposed approach is shown in Figure 1.

3.1. Social Science and Psychology Features

Existing research has used various ad-hoc feature sets from rumor data for the rumor veracity classification task. The classification result is reported based on machine learning techniques without considering social science and psychology theories. Little work has tried to understand and build the feature sets from grounded theoretical work about rumor spread from social science and psychology. Linguistic Inquiry and Word Count to build a set of features that reflect the process of doubt, negation, and guessing of rumor propagation were used in [11, 25]. In contrast, we adopt two rumor theories that describe how user decision making process affects the spread of a rumor and its veracity. The first is the rumor spread theory "The Basic Law of Rumor" [9] where the truth and strength of a rumor depend on the importance of its topics and the significance of its ambiguity. For example, if a rumor is about an important individual (e.g., "Obama is a Muslim"), it is more likely to be spread further and more likely to be false. In addition, rumors that are hard to verify will likely last longer. Researchers have shown that the more controversial the comments, the more popular the post will be [26]. Intuitively, we take advantage of extracted Wikipedia topics from rumor data to represent the importance of a rumor (see Section 3.1) and the sentiment score of users' comments to represent the controversy of a rumor (see Section 3.2).

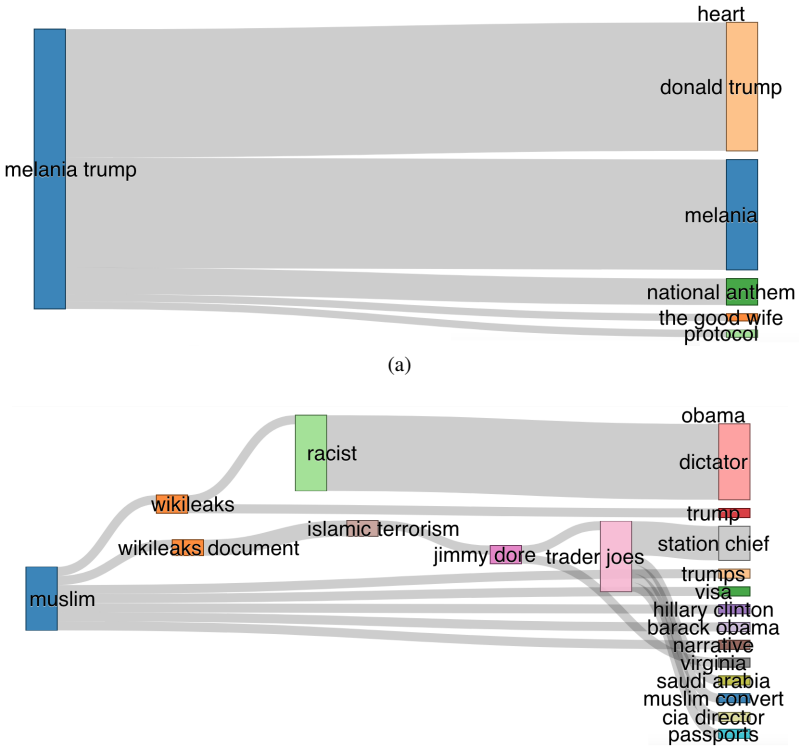


Figure 2: Topic features between true vs. false rumors: a) "Melania Trump Reminds the President to Put His Hand Over His Heart", b) "Obama is a Muslim". In this graph, each node is a Wikipedia topic that is extracted from the submission title. There is a connection between two continuous submissions of two topics if their semantic similarity score is above 0.5. For the true rumor, we have a concise list of topics, while we have a wide range of topics about the false rumor "Obama is a Muslim". The node and connection size and color are generated automatically based on the number of topics for the best visual layout.

The second is the rumor spread model theory of Daley and Kendall [27]. We compute the numbers of *spreaders*, *ignorants*, and *stiflers* in a rumor throughout its life cycle as discussed in [28]. In this model, N is the number of users who interact with this rumor. In the beginning, one user learns about this rumor from another source and tries to spread it by posting a submission. Other users will read this submission and start spreading to other members. In each submission, a user is categorized into one of the three categories: spreaders, ignorants, and stiflers, which are denoted as S, I, and R, respectively. Spreaders are people who actively spread the rumor; Ignorants are people who are ignorant of the rumor at first but will become either spreaders or stiflers at a later stage; Stiflers are people who posted about the rumor, but are no longer interested in spreading it. This rumor spread model is modeled as a stochastic process on $P = \{S, I, R\}^N$, where N is the total number of users for a rumor in the dataset. Let the state of user i at time t be a function of time $X_i(t)$. The procedure to compute $X_i(t)$ is as follows:

- The user who posted the first submission about this rumor at time $t = 0$ is the spreader ($|S| = 1, |I| = N - 1, |R| = 0$).
- Users who reply to the first submission are ignorants at t_0 and will become spreaders at t_1.
- At time t_i:
  - If user j posts this submission, user j is a spreader.
  - If user j has a comment on this submission:
    - User j is a spreader if user j has a comment at t_{i-1}.
    - User j is a stifler if user j has a comment at time from t_{i-2} to t_0.
    - User j is an ignorant if user j has no comments at time from t_{i-1} to t_0.
    - User j will become a spreader at time t_{i+1}.

Based on those two rumor spread theories, we build various features for the rumor veracity classification task.

3.2. Topic Features

Previous studies highlight the importance of topics that affect the popularity of a rumor. The importance of a topic plays a significant role in the popularity of rumor spread according to [29], while users spread rumors when they feel anxious about a topic they are interested in (e.g., AIDS-related rumors) according to [30]. For each submission, we use the Dexter topic extraction tool, [31] to obtain all Wikipedia topics in submission titles. We use the number of topics in each rumor to determine how important this rumor is. Also, we compute the approximate entropy (ApEn) for each topic list of a rumor. ApEn is a method to evaluate the regularity and the unpredictability of the fluctuating nature of temporal data. Researchers have used this approach to compute topic evolution models [32, 33]. Our first assumption is that true rumors are usually verified in a short time, while false rumors take longer to be debunked. The second assumption is that the topics of true rumors are more regular and predictable than false rumors. An example of topic evolution between a true rumor and a false rumor is shown in Figure 2. We observe that the topics of true rumors are more regular and less fluctuating than false rumors.

For each rumor RU_i, we have a list of topics $T_1, T_2, ...T_n$ in ascending order. To calculate the distance between two topics for ApEn, we compute the semantic similarity between two topics, T_1 and T_2, using Google Tri-gram Method [34].

3.3. Sentiment Features

Online rumors could draw attention, stimulate involvement, and influence attitudes and actions of OSN users [35]. In an online conversation, user sentiment significantly contributes to how news or topics become popular [26]. Each user comment is parsed into sentences and each sentence is assigned a sentiment score: "Positive","Negative", or "Neutral" using the OpenNLP toolkit. We apply the concepts of sentiment polarity and subjectivity of Zhang and Skiena [36] for each rumor in our ground-truth dataset as follows:

$polarity\_score = (p - n)/(p + n)$

$subjectivity\_score = (n + p)/N$

where p is the number of positive statements, n is the number of negative statements, and N is the total number of statements (including neutral statements). Polarity score represents whether a rumor is associated with the entity positively or negatively, while subjectivity score depicts how much sentiment a rumor garners.

3.4. Network Structural Features

The questions of who spreads rumors and how have been studied extensively in the literature [3, 6]. Researchers have stated that rumors are usually spread by few influencing users, and these users could spread rumors a lot quicker and cause significant damage to the individual targets in OSNs [12]. Using this assumption, we compute the betweenness and closeness of the user interaction directed graph as shown in Figure 3 where a node is a user, and an edge between two nodes represents that a user (denoted by one node) replies to a comment of another user (denoted by another node). A user with a high betweenness centrality score could propagate rumorous news to a large user population in the network and influence the popularity of a rumor, while a user with a high closeness centrality score could transmit the rumor to a large user population in a short time. These users play a crucial role in detecting the truth of rumors in its early stage. As influencing users could significantly impact the popularity of a rumor and its truthfulness, we use the highest betweenness and closeness scores in the user graph as features for the classification task.

3.5. Content Analysis - the Wisdom of the Crowd

Table 1: Feature sets that are established from ground-up social science theories and other Basic Features (BF). All features are derived from Section 3.

Group	Features
Social Science and Psychology	Number of spreaders
	Number of stiflers
	Number of ignorants
Topic	Number of topics
	Approximate entropy score
Sentiment	Polarity score
	Subjectivity score
Network Structural	Number of submissions (BF)
	Number of comments (BF)
	Number of unique users (BF)
	Betweenness centrality score
	Closeness centrality score
Content	TF-IDF user comment vectors

Previous studies have used various features to distinguish between rumors and non-rumors [19, 11].

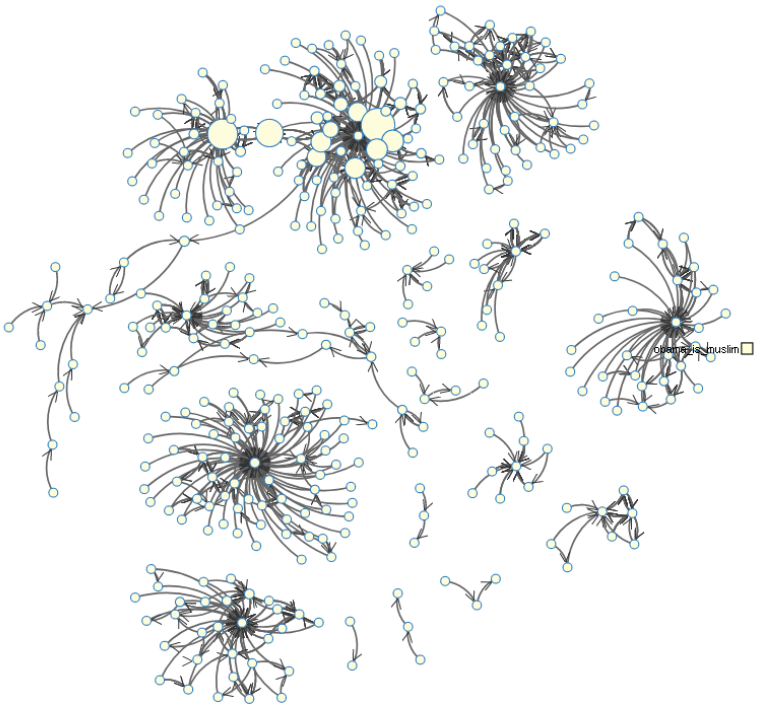


Figure 3: User interaction directed graph of the false rumor "Obama is a Muslim". Each node is a user and an edge represents the connection between two users (who replies to who). The node size represents the centrality score of the users.

In this paper, we study if we could identify the level of truth of rumors based on how users respond to these rumors. Why people spread rumors in a social network psychologically has been studied [30]. People spread rumors based on three motivations: (1) fact-finding, (2) relationship-building, and (3) self-enhancement. Fact-finding people intend to find the truth of rumors through a problem-solving process. In contrast, those motivated by relationship-building are simply interested in communicating and interacting with other individuals by sharing information about particular rumors. Self-enhancement people are either consciously or unconsciously approving false rumors. How users interact with each other within a rumor was studied in [6], finding that a dominant number of users just try to joke about this rumor. Also, there are more users who try to disapprove a false rumor than to approve it. Based on this finding, we aim to identify whether we could use the wisdom of the crowd (i.e., users' comments) to debunk a rumor and find its truth. Social science and Psychology, Topic, Sentiment, Network Structural, and Content features for the rumor veracity classification task are shown in Table 1.

This section describes the analysis of the data and highlights the characteristics of rumors in each category using the experimental results. We first report the results on how to accurately classify rumors into one of the five categories. Next, we treat the rumors in the "Mostly False" to be the same as the category "False", and the rumors in the "Mostly True" to be the same as the category "True". This combination results in a three-class classification task: "False", "True", and "Half True". Finally, we filter out the rumors in the category "Half True" and report the result for a two-class classification task.

3.6. The Newly-created Reddit Rumor Dataset

For each rumor, we need to collect the following elements: 1) The truth about this rumor, 2) Posts (data) about this rumor, and 3) Metadata about this rumor, such as sentiment analysis, topics, and user interaction graphs. For the dataset, we first identified the most popular rumors (58 true and false rumors) from previous work [11, 10]. In addition, we collect 30 new rumors from Snopes.com and Politifact.com that could belong to one of the three new categories: "Mostly False", "Mostly True", and "Half True". The labels of the combined rumor list are verified with the five categories from these websites. We identified a total of 88 rumors (see

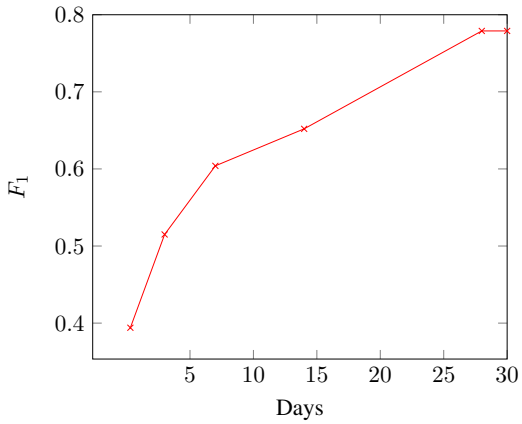


Figure 4: Five-class classification results over time. The system achieves the best F_1 score (0.78) on day 28.

the Supplementary Material - Rumor Description) and extracted submissions if a submission contains explicit keywords relevant to the rumor. We adopt RumourFlow [28] to collect and visualize rumor data in Reddit. All collected rumor data are provided to the visualization system through a JSON restful web service and a JAVA backend. The goal of using RumourFlow visualization is to analyze and understand the characteristics of rumors in each category and confirm whether the feature sets derived from social science work could be applied.

Overall, we collected 88 rumors that belong to one of the five categories: "False", "Mostly False", "True", "Mostly True", and "Half True". For each rumor category, we also report the number of long-standing rumors vs. breaking news rumors. Long-standing rumors are rumors that have been discussed and propagated for a long period while breaking news rumors are usually circulated in breaking news events, such as natural disaster, political events in their early circulation [24]. The long-standing rumors are dominant in the dataset. A detailed statistics of the newly-created rumor dataset is shown in Table 2. We observed that false rumors and mostly false rumors receive the highest number of discussed submissions and comments. This supports the assumption that false rumors are more popular than true rumors [12].

Table 2: The newly-created Reddit rumor ground-truth dataset. The table also shows the number of long-standing rumors in each category.

Category	No. Rumors	Avg. No. Submissions	Avg. No. Comments	Long-Standing Rumors
False	48	14	249	34
Mostly False	10	11	198	7
True	10	8	98	7
Mostly True	10	87	8	8
Half True	10	6	99	8

Table 3: Classification results for different sets of categories.

Category	Accuracy	Recall	F_1
Five classes	0.589	0.545	0.564
Three classes	0.670	0.670	0.670
Two classes	0.752	0.744	0.747

4. Results

4.1. Content Feature Classification

We transform users' comments on rumors into TF-IDF vectors, the elements of which reflect how important a word is in a document or in a corpus (stop words are removed). Each rumor is represented by a vector of the 200 highest ranking words. We choose Naive Bayes classification (NB) with ten-fold cross validation from Weka [37], as NB is fast to build and could be trained with less data. Those two characteristics of NB are important for the rumor veracity task. The results of the classification are shown in Table 3.

Table 4: Classification results for topic, sentiment and structural features.

Category	Accuracy	Recall	F_1
Five classes	0.484	0.398	0.419
Three classes	0.631	0.534	0.540
Two classes	0.781	0.628	0.650

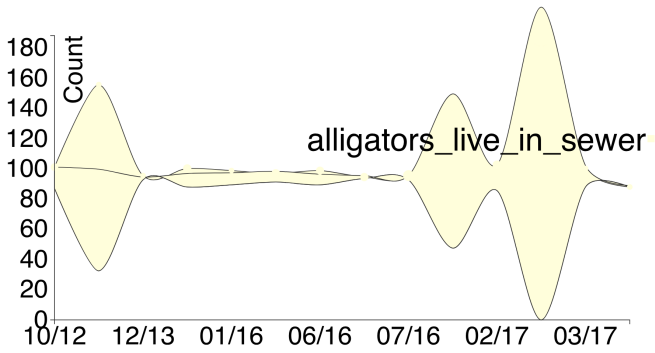
Table 5: Classification results for topic, sentiment and structural features.

Category	Accuracy	Recall	F_1
Five classes (5C)	0.795	0.729	0.761
5C minus structural	0.593	0.729	0.654
Three classes (3C)	0.864	0.852	0.856
3C minus minus structural	0.745	0.792	0.768
Two classes (2C)	0.927	0.879	0.903
2C minus structural	0.795	0.729	0.761

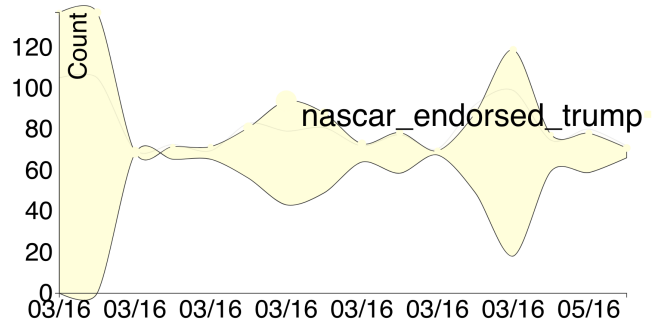
For the five class result, we achieve 56.4% F_1 score. The results get better with three classes and two classes with 67% and 74.7% F_1 score respectively. The result shows that the proposed approach can better distinguish the difference between true rumors and false rumors. We achieve the best result without the category "Half True". This is because "Half True" rumors are very controversial and not easily identifiable.

4.2. Social Science and Psychology, Topic, Sentiment, and Network Structural Feature Classification Result

We report the classification results using social science and psychology, topic, sentiment, and network structural features in Table 4. We have the same pattern as



(a)



(b)

Figure 5: Long-standing vs. breaking news rumors a) The long-standing false rumor "(alligators live in sewers) " has various peaks and the highest peak is usually not the first peak, b) The breaking news rumor "Nascar endorsed Trump" has the highest peak first and smaller peaks occurred later. The x-axis represents the time, while the y-axis represents the number of comments for each submission.

using the content features, but the results do not perform as well as using the content features. The classification results of two and three classes are better than the five classes in terms of accuracy and recall. We again observe that the classification results are significantly better for the two-class classification task. This supports that the rumors in the category "Half True" are harder to detect. Although the system achieved lower accuracy than the content features, the results are still close. We plan to integrate the intrinsic characteristics of the two feature sets aiming to achieve a better result than either.

4.3. Combined Feature Classification Result

After combining the social science and psychology, sentiment, topic, and network structural features with the 200 attributes of content features, we achieve the best result as shown in Table 5. We achieve the best precision and recall using only two classes, and this result is comparable with the two-class veracity classification results reported in [11] on a different dataset. We also evaluate the importance of the three social science features: number of spreaders, number of stiflers, and number of ignorants from the social science and psychology group by performing the ablation test. The three attributes that are built from Daley and Kendall's stochastic rumor spread theory are demonstrated to significantly improve classification quality, as shown in Table 5.

4.4. Rumor Truth Time-varying Result

We also investigate whether we can detect a rumor in its earlier stages and still maintain accuracy. As rumors may have different peak cycles, we build the combination feature sets based on different time windows and classify using the following intervals: three hours, three days, seven days, 14 days, and 28 days. Classification achieves an F_1 score above 60% after the first seven days and the result after 28 days is comparable to the best result in Table 5, as shown in Figure 4. The system did not perform well when trying to find the level of truth of rumors after three hours or three days.

We observe and compare the fluctuating nature between long-standing vs. breaking news rumors using RumourFlow in Figure 5. The long-standing rumors

have various peaks over a long period, and the highest peak of comments is usually reached after the rumor has been discussed for a while. On the other hand, the breaking news rumors usually have the highest peak at the beginning and several smaller peaks until they die. We select rumor data from the beginning until the highest peak of comments. For the first peak, we select rumor data from the beginning until the first peak of comments. Using this finding, we investigate the performance of the veracity classification task after a rumor's first peak and highest peak. For the five-class veracity classification task, we achieve the best $F_1 = 0.78$ using rumors' highest peak and $F_1 = 0.73$ using the rumors' first peak. This is an important finding as the system could effectively detect the level of truth for breaking news rumors in a short time (the average highest peak of breaking news rumors in the dataset is three hours).

5. Discussion and Conclusion

In this paper, we introduce a new Reddit rumor dataset where each rumor is categorized into one of five categories: "False", "Mostly False", "True", "Mostly True", and "Half True". The truth levels of rumors in this dataset better reflect the fine-tuned truth values of rumors in Snopes.com and Politifact.com. Next, we investigate whether the proposed approach can effectively detect and debunk the truthfulness of rumors through an extensive set of features including social science and psychology theories. The experiments show that our system can efficiently detect the truthfulness of rumors. This could bridge the gap between social science theories and experimental research of rumor spread in online social networks.

We also explore various feature sets and levels of rumor veracity in the experiments. We found that the system best detects rumors in two classes "False" or "True". Our two-class rumor veracity classification result is comparable with the the-state-of-the-art method in the literature [11]. The "Half True" category degrades the classification result. One of the underlying reasons is the conflicting characteristics of such rumors. We also found similar characteristics between "False" and "Mostly False" as well as between "True" and "Mostly True" rumor categories. This shows that social media users do not distinguish between those two categories. Fundamental future work is to study the underlying similarities and differences between those two pairs of rumor categories ("Mostly False" vs. "False" and "Mostly True" vs. "True") so that the system could better distinguish them. With the five levels of rumor truthfulness, the newly-created rumor dataset provides a better rumor truth mapping with existing fact-checking websites (e.g., Snopes.com and Politifact.com). In addition, solving the problem of the "Half True" rumors is an urgent need as it is more and more popular in political news.

Early detection of the truth of rumors is a key factor in preventing their spread. The experiments show that we could effectively debunk rumors as early as in seven days. We also find that the proposed approach can efficiently find the truth of rumors after their first peak ($F_1 = 0.73$). Hence, it is possible to effectively detect the truth levels of breaking news rumors within three hours. On the other hand, the long-standing rumors could be efficiently debunked after the rumors' highest peaks (usually after 28 days). Due to Twitter API limits and lack of availability of sufficient data in the existing datasets in the literature, we have not been able to construct and evaluate our proposed feature set derived from social sciences and psychology with those datasets. Important future work will be to extend and compare the results of the proposed Reddit rumor dataset with other comparable and available rumor datasets (e.g., rumors on Twitter).

With the combination of established social science theories, visualization, and advanced machine learning techniques, the developed system could effectively detect the rumor truthfulness in its early stage. This could bring benefits to both practical applications and theoretical research. For example, the system could be used by social media websites (e.g., Facebook, Twitter, or Reddit) and fact-checking websites (e.g., Snopes.com and Politifact.com) to filter out and limit the spread of false rumors or fake news in their early stage. This helps OSN users to have an educated decision-making process. On the other hand, the results could validate and confirm established social science theories to a certain extent. Essential future work is to implement a controlled experiment to see how effective the developed system could influence OSNs user decision-making process in a rumor (spread or debunk a rumor).

6. Acknowledgement

The research was funded in part by ACENET in Atlantic Canada (https://www.ace-net.ca/), Compute Canada (www.computecanada.ca), Natural Sciences and Engineering Research Council of Canada, and International Development Research Centre, Ottawa, Canada.

References

[1] R. L. Rosnow, "Inside rumor: A personal journey.," *American Psychologist*, vol. 46, no. 5, p. 484, 1991.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *WWW*,

pp. 591–600, ACM, 2010.

[3] B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Communications of the ACM*, vol. 55, no. 6, pp. 70–75, 2012.

[4] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, "Theory of rumour spreading in complex social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457–470, 2007.

[5] D. Shah and T. Zaman, "Rumors in a network: who's the culprit?," *Information Theory, IEEE Transactions*, vol. 57, no. 8, pp. 5163–5181, 2011.

[6] A. Dang, M. Smit, A. Moh'd, R. Minghim, and E. Milios, "Toward understanding how users respond to rumours in social media," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 777–784, IEEE, 2016.

[7] M. Lukasik, P. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn, "Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 393–398, 2016.

[8] L. Festinger, D. Cartwright, K. Barber, J. Fleischl, J. Gottsdanker, A. Keysen, and G. Leavitt, "A study of a rumor: its origin and spread.," *Human Relations*, 1948.

[9] A. Chorus, "The basic law of rumor.," *The Journal of Abnormal and Social Psychology*, vol. 48, no. 2, p. 313, 1953.

[10] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," *arXiv preprint arXiv:1704.05972*, 2017.

[11] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PLOS ONE*, vol. 12, no. 1, p. e0168344, 2017.

[12] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?," in *Proc. of the first workshop on social media analytics*, pp. 71–79, ACM, 2010.

[13] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," tech. rep., National Bureau of Economic Research, 2017.

[14] N. DiFonzo and P. Bordia, "Rumor and prediction: Making sense (but losing dollars) in the stock market," *Organizational Behavior and Human Decision Processes*, vol. 71, no. 3, pp. 329–353, 1997.

[15] Snopes, "Snopes." `http://www.snopes.com`, 2017. [Online; accessed 19-July-2017].

[16] Politifact, "Politifact." `http://www.politifact.com/`, 2017. [Online; accessed 19-July-2017].

[17] V. Ambati, S. Vogel, and J. G. Carbonell, "Active learning and crowd-sourcing for machine translation," in *LREC*, 2010.

[18] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proc. of the ACL 2012 System Demonstrations*, pp. 115–120, Association for Computational Linguistics, 2012.

[19] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *EMNLP*, pp. 1589–1599, Association for Computational Linguistics, 2011.

[20] G. Giasemidis, C. Singleton, I. Agrafiotis, J. R. Nurse, A. Pilgrim, C. Willis, and D. V. Greetham, "Determining the veracity of rumours on twitter," in *International Conference on Social Informatics*, pp. 185–205, Springer, 2016.

[21] L. Derczynski and K. Bontcheva, "Pheme: Veracity in digital social networks.," in *UMAP Workshops*, 2014.

[22] P. Lendvai and U. D. Reichel, "Contradiction detection for rumorous claims," *arXiv preprint arXiv:1611.02588*, 2016.

[23] A. Zubiaga and H. Ji, "Tweet, but verify: epistemic study of information verification on twitter," *SNAM*, vol. 4, no. 1, p. 163, 2014.

[24] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 32, 2018.

[25] N. DiFonzo and P. Bordia, "Corporate rumor activity, belief and accuracy," *Public Relations Review*, vol. 28, no. 1, pp. 1–19, 2002.

[26] L. Dang-Xuan and S. Stieglitz, "Impact and diffusion of sentiment in political communication-an empirical analysis of political weblogs.," in *ICWSM*, 2012.

[27] D. J. Daley and D. G. Kendall, "Stochastic rumours," *IMA Journal of Applied Mathematics*, vol. 1, no. 1, pp. 42–55, 1965.

[28] A. Dang, A. Moh'd, E. Milios, and R. Minghim, "What is in a rumour: Combined visual analysis of rumour flow and user activity," in *CGI*, pp. 17–20, ACM, 2016.

[29] R. L. Rosnow and E. K. Foster, "Rumor and gossip research," *Psychological Science Agenda*, vol. 19, no. 4, pp. 1–2, 2005.

[30] P. Bordia and N. DiFonzo, "Psychological motivations in rumor spread," *Rumor Mills: The Social Impact of Rumor and Legend*, pp. 87–101, 2005.

[31] S. Trani, D. Ceccarelli, C. Lucchese, S. Orlando, and R. Perego, "Dexter 2.0: an open source tool for semantically enriching data," in *Proc. of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pp. 417–420, CEUR-WS. org, 2014.

[32] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. of the 23rd international conference on Machine learning*, pp. 113–120, ACM, 2006.

[33] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, pp. 121–128, 2008.

[34] A. Islam, E. E. Milios, and V. Keselj, "Text similarity using google tri-grams.," in *Canadian Conference on AI*, vol. 7310, pp. 312–317, Springer, 2012.

[35] N. DiFonzo and P. Bordia, *Rumor psychology: Social and organizational approaches.* American Psychological Association, 2007.

[36] W. Zhang and S. Skiena, "Trading strategies to exploit blog and news sentiment.," in *ICWSM*, 2010.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.