A Comparative Study of Transactional and Semantic Approaches for Predicting Cascades on Twitter

Yunwei Zhao[1], **Can Wang**[2,*], **Chi-Hung Chi**[3], **Kwok-Yan Lam**[4], **Sen Wang**[2]
[1] The National Computer Network Emergency Response Technical Team/Coordination Center of China
[2] School of Information and Communication Technology, Griffith University, Australia
[3] Data 61, CSIRO, Australia
[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore
zhaoyw@cert.org.cn, can.wang@griffith.edu.au, chihungchi@gmail.com,
kwokyan.lam@ntu.edu.sg, sen.wang@griffith.edu.au

Abstract

The availability of massive social media data has enabled the prediction of people's future behavioral trends at an unprecedented large scale. Information cascades study on Twitter has been an integral part of behavior analysis. A number of methods based on the transactional features (such as keyword frequency) and the semantic features (such as sentiment) have been proposed to predict the future cascading trends. However, an in-depth understanding of the pros and cons of semantic and transactional models is lacking. This paper conducts a comparative study of both approaches in predicting information diffusion with three mechanisms: retweet cascade, url cascade, and hashtag cascade. Experiments on Twitter data show that the semantic model outperforms the transactional model, if the exterior pattern is less directly observable (i.e. hashtag cascade). When it becomes more directly observable (i.e. retweet and url cascades), the semantic method yet delivers approximate accuracy (i.e. url cascade) or even worse accuracy (i.e. retweet cascade). Further, we demonstrate that the transactional and semantic models are not independent, and the performance gets greatly enhanced when combining both.

1 Introduction

Predicting information cascades [Taxidou and Fischer, 2014; Hoang and Mothe, 2017] has been an integral part of the behavior analysis, since the future popularity trend of events indicates the intensity with which people would react, and hence will have a great and direct impact on decision makings. A cascade of information on Twitter refers to a collection of tweets reproduced from a specific piece of information, of which the most three common forms include a hashtag, a url and a tweet. The corresponding cascades are called the hashtag cascade, the url cascade, and the retweet cascade. Among those cascade mechanisms, the hashtag cascade and the url cascade are based on the clickable links in the content. A hashtag is a concise and accurate content descriptor, expressed as # followed by a word, which takes into account the human perceptions of the content. A url is a hyperlink that directs to external text. The retweet cascade, on the other hand, is a reproduction of the content itself. It includes both the direct retweet (i.e. DT for short) via clicking the retweet button and the modified retweet (i.e. MT for short) in the format of "RT @ [the users you give credit to] [original content]". These diffused topics are usually taken as a central starting point from which to investigate a number of socio-economic phenomena, e.g. stock market trend or the potential success of a candidate in a political election. The analysis and prediction of information cascades have become more important in the last decade due to the wide use of social media/networks.

Despite the large number of users actively participating in online social media, and the ever increasing amount of information available, people still face the challenges to properly characterize the diffused cascades to make predictions about the future trends of the observed cascades. A number of methods have been suggested, and two broad categories can be identified in the literatures: (i) transactional approach and (ii) semantic approach. The former approach includes the term frequency-inverse document frequency (i.e. TF-IDF for short) model [Benhardus and Kalita, 2013] and the latent Dirichlet allocation (i.e. LDA for short) model [Blei *et al.*, 2003; Hong *et al.*, 2011]. The most notable example of the semantic approach includes the Google-profile of mood states that measure the mood in terms of six dimensions: calm, alert, sure, vital, kind, and happy [Bollen *et al.*, 2011]. Other examples are based on the controversy, the hotness/virality [Guerini *et al.*, 2011], and etc.

However, these two approaches are usually investigated separately. An in-depth understanding of the pros and cons of the transactional based and the semantic based models is lacking. It is unclear to both the researchers and industry practitioners how the transactional driven approach differs from

*Corresponding author: Can Wang (can.wang@griffith.edu.au)

the semantic driven diffusion prediction method, and how the difference in diffusion mechanisms affects the prediction performance. Hence, we are highly motivated to work on the research questions being addressed in this paper. How do these two approaches differ under different diffusion mechanisms? Under what conditions one approach outperforms the other? and vice versa. Furthermore, are these two approaches independent, or is there any overlapping in predicting cascades and at what degree? Will it be more effective if we combine different methods together?

In this paper, we conduct a comparative study of the transactional approach and the semantic approach in predicting whether a cascade of information will continue to diffuse in the future, with the Twitter data from 2010.01 to 2010.10. For the transactional model, we restrict our attention to those two most widely used methods in the feature extraction: (i) TF-IDF [Benhardus and Kalita, 2013], and (ii) a dynamic variant of LDA model– Dynamic Topic Models (DTM for short) [Blei and Lafferty, 2006]. The semantic features of a cascade are described from the following five aspects: sentiment, controversy, content richness, hotness, and trend momentum. The contributions of this paper are listed as follows:

- We compare the pros and cons of the semantic and the transactional models. The results show the effectiveness of both approaches differs with respect to cascade mechanisms. In the cascade with the higher content complexity, i.e. hashtag cascade, the prediction accruacy of semantic approach is higher, indicating that the semantic model provides a better way of feature representation. In the url cascade and the retweet cascade with more explicit and directly observable patterns in an individual tweet, thus with the lower content complexity, the transactional approach leads to the better prediction accuracy.
- We analyze the inter-relations between semantic and the transactional models, and find out their overlapping degree is above 80% for hashtag cascade and url cacade.
- We empirically study Twitter data, on which the performance of semantic and transactional approaches is correlated with the content complexity in different cascade mechanisms. A great enhancement in the prediction accuracy can be achieved when combining both.

2 Related Work

The emergence of an extensive focus on human behaviors in information cascade has started with the advent of social media [An *et al.*, 2014]. Two broad categories are identified:

The *transactional features* that predict the information cascading behavior include: (i) network transmission statistics [Hong *et al.*, 2011]; (ii) user behavior statistics [Takahashi *et al.*, 2011]; (iii) text-based statistics [Akbari *et al.*, 2016; Cheng *et al.*, 2014; Hong *et al.*, 2011]. The most widely used text-based feature extraction method is the TF-IDF model [Benhardus and Kalita, 2013]. LDA model [Hong *et al.*, 2011] extracts a topic distribution feature for each document. Other models incorporate the social properties, such as Taxidou's work [Taxidou and Fischer, 2014] that considers the most recent, the most retweeted, and the most followed influencer; and the dynamic community topic model [An *et al.*, 2014] that captures the dynamic features of communities and topics. When the dimensionality of the features extracted is large, feature reduction methods are usually adopted [Wang *et al.*, 2014b; 2015].

The *semantic features* are commonly defined to explain user behaviors, such as public affair attitude tracking [An *et al.*, 2014], stock market trend forecasting [Bollen *et al.*, 2011; Sprenger *et al.*, 2013]. Notable examples include the sentiment: the polarity of affection (i.e. positive vs. negative). For instance, a correlation between the short-term fluctuations in negative sentiments and the major climate events was explored in [An *et al.*, 2014]. The sentiment classification task is still challenging up-to-date, linguistic based psychometric tools [Bradley *et al.*, 1999; Bollen *et al.*, 2011] are most widely adopted, other methods include Recurrent Random Walk Network models [Zhao *et al.*, 2017]. Controversy [Chew and Eysenbach, 2010; Guerini *et al.*, 2011], or referred to as "disagreement in the content", is used in stock market trending analysis [Sprenger *et al.*, 2013]. Bollen et al. [Bollen *et al.*, 2011] proposed the Google-profile of mood states via calm, alert, sure, vital, kind, and happy.

There are several model-based methods [Zhao *et al.*, 2015; Kobayashi and Lambiotte, 2016] where the number of retweets is modelled as a self-exciting point process. Even though the differences between users are considered by explicitly taking the behavior characteristics into consideration, however, it is on a exterior statistic basis. Thus, it is out of the scope of this paper. It is expected that in future work, by introducing an intermediate layer of the behaviour interior dimensions, the interpretation of the raw data in the dynamic diffusion process through the model-based methods is to be greatly enhanced and improved.

3 Transactional vs. Semantic

In this section, we present an overview of the transactional and semantic approaches, and identify their similarity as well as difference. It provides the required background for this work and motivates our evaluation methodology.

3.1 Transactional Approach

The TF-IDF and LDA models are the most popular methods that are based on transactional statistics (i.e. word frequency).

The TF-IDF model [Benhardus and Kalita, 2013] is a widely used method for feature extraction in the application of information retrieval and text mining. For a given document (e.g. the cascaded information in the context of this paper), a feature vector with the TF-IDF metric for each word is extracted. The TF-IDF metric contains two statistical indexes: term frequency and inverse document frequent, measuring both locally and globally how important a word is to a document, respectively. It therefore provides a direct visual representation of the key words that are unique to the document, through which it establishes the correlation between different cascaded information and thereby performs the categorization and prediction tasks.

The LDA model [Blei *et al.*, 2003; Hong *et al.*, 2011] reduces the high dimensionality of TF-IDF model by extracting a topic distribution feature for each document. Simply put,

the LDA is a generative model of a corpus, where each document is a random mixture over the latent topics, and each topic is regarded as a distribution over words. However, the topics obtained in each time slot are independent from each other. Hence, we resort to the dynamic topic model (i.e. DTM for short) [Blei and Lafferty, 2006] which considers the time evolution of topics, and make the topic features of each document (or cascade) comparable to one another.

3.2 Semantic Approach

The semantic features that disclose interior driving forces and intrinsic cause-effects of the cascades [Wang *et al.*, 2014a] are summarized into three aspects as below.

Sentiment and Controversy

Sentiment refers to the affection valence (i.e. positive vs. negative), and controversy is about the degree to which people occupy different stances towards certain contents [Guerini *et al.*, 2011]. Linguistic based psychometric tools [Bradley *et al.*, 1999] are usually used to measure the sentiment, as given in Equation (1),

$$\text{Sent}_i^t = avg(\text{Sent}(p_i^{j,t})), \tag{1}$$

where $Sent(p_i^{j,t})$ is the sentiment of post j in the i-th cascade within the t-th week, measured by the word affect valence based on the ANEW averaged by word appearance frequency.

The controversy is approximately measured by calculating the average sentiment differences between two consecutive posts of the i-th cascade, as given in Equation (2).

$$\text{Contro}(t,i) = avg(|\text{Diff}(\text{Sent}(p_j,i), \text{Sent}(p_{j-1},i)|). \tag{2}$$

It reveals that the sentiment describes the favorite tendency towards the content, i.e. "attractiveness" (positive valence) or "aversiveness" (negative valence). The controversy provides an approximation of the stances that people hold towards certain contents [Guerini *et al.*, 2011], i.e. for and against.

Content Richness

Content richness refers to the diversification degree of the content spread within a specific topic. It has two sub-dimensions: content volume and content intensity. Metrics such as the average content length [Naaman *et al.*, 2011], the entropy [Tononi and Sporns, 2003], and the string metrics [Masucci *et al.*, 2011a; 2011b], are usually applied to measure these two sub-dimensions. The Cascade i's content richness at the time t is a weighted sum of the normalized unique content length and the content intensity, see Equation (3) :

$$\begin{aligned}\text{CR}(t,i) = \alpha \cdot \text{uwl}(t,i) + \beta \cdot 2/(N_t^i(N_t^i-1)) \\ \sum_{j_2=1\ldots N_t^i} \sum_{j_1<j_2} \frac{\text{ld}(p_{j_1},p_{j_2})}{\max(\text{length}(p_{j_1}),\text{length}(p_{j_2}))},\end{aligned} \tag{3}$$

where $uwl(t,i)$ is the normalized unique word length of the tweets within the i-th period, here, we normalize it to [0,1], $\text{ld}(p_{j1}, p_{j2})$ is the Levenshtein Distance between any two posts p_{j_1} and p_{j_2}, N_t^i is the post amount within topic t at time i, and $\alpha + \beta = 1$. The content richness simply equates the content volume when α = 1. We have $\text{uwl}_i^t \in$ [0,1], $\text{lev}_i^t \in$ [0,1], $\text{CR}_i^t \in$ [0,1]. In our experiment, we set α = 0.5, β = 0.5.

Hotness and Trend Momentum

Hotness refers to the intense and immediate focus of users on a specific cascade, while trend momentum describes the tendency of cascade to spread quickly in the community. Therefore, the measurement of these two sematic features includes two aspects: the communication count and the coverage-of-people. Accordingly, the cascade i's hotness and trend momentum at time t can be given by Equations (4) and (5):

$$H(t,i) = \alpha \cdot \text{CC}(t,i) + \beta \cdot \text{CovP}(t,i), \tag{4}$$

$$\text{TM}(t,i) = \alpha \cdot \text{TMCC}(t,i) + \beta \cdot \text{TMCovP}(t,i), \tag{5}$$

where $\text{CC}(t,i)$ and $\text{CovP}(t,i)$ are the normalized communication count and the normalized people coverage of topic t at time i, respectively. $\text{TMCC}(t,i) = |\text{CC}(t,i)–\text{CC}(t,i-1)|$ and $\text{TMCovP}(t,i) = |\text{CovP}(t,i)–\text{CovP}(t,i-1)|$ denote the corresponding trend momentum of communication count and coverage of people, respectively, and $\alpha + \beta = 1$. $H(t,i)$ simply equates the traditional diffusion degrees when α = 1.

	n_w		
	Hashtag Cascade	Url Cascade	Retweet Cascade
w = 1	276910	229972	124764
w = 2	32143	86970	18047
w = 3	12489	18428	3682
w = 4	7010	7588	1421
w = 5	4430	3911	698
w = 6	3053	2378	389
w = 7	2159	1565	244
w = 8	1685	1088	154

Table 1: Statistics of Data

4 Evaluation Methodology

We discuss the methodology we use to evaluate the similarity and difference of the semantic approach and the transactional approach from (i) the prediction accuracy, and (ii) the interrelations between both approaches.

We collect 10-month data from Twitter of about 112,044 users with around 78 million tweets, from the 2nd week to the 42nd week in 2010. The tweets are collected through crawling the followers' and followees' tweets of a random selection of active users, with the number of hashtags more than 1000. The sentiments of the tweets are evaluated based on ANEW [Bradley *et al.*, 1999] with emoticons [1] and slangs[2] replaced with the corresponding text. The data is filtered on the basis of whether users are English-speaking users, as most psycholinguistic tools such as ANEW are only concerned with English words. The statistics is shown in Table 1, where n_w denotes the number of cascaded threads with the time length w weeks (i.e., a cascade has posts for at least w weeks).

Prediction Accuracy Comparison

The cascade prediction problem is a binary classification task. That is to say, if the topic in the test set diffuses, the desired output label λ is "True", otherwise is "False". Based on the extracted features, no matter transactional features or semantic features, we apply the KNN [Bai *et al.*, 2015] to find the nearest k (k=5) neighbors and make the prediction based on the majority class labels. The accuracy is measured by precision and recall [Wang and Wang, 2007].

Three cascade mechanisms are investigated here, namely, the hashtag cascade, the url cascade, and the retweet cascade. The inherent difference in the cascade mechanisms decides that the content compositions in these three ways are distinct.

[1] http://cool-smileys.com/text-emoticons, with 938 emoticons

[2] http://www.noslang.com/, with 5396 slangs and abbreviations

In the url cascade and the rewteet cascade, the proportion of the words in a tweet that get reproduced from the source tweet takes up a higher percentage, while it is only a very small percentage in the hashtag cascade as the reproduced source information is only the keywords marked with "#". This difference in cascade mechanisms may lead to discrepancy in the prediction performance. For the transactional approach, especially TF-IDF, is more directly based on the keywords. While for the semantic approach, though some of its metrics are linguistic, e.g. sentiment and content richness, it has stronger descriptive capability in the context of diverse contents.

We argue that the effectiveness of transactional approach and semantic approach is correlated with the content complexity. More specifically, the transactional approach has better prediction accuracy in the url cascade and the retweet cascade where the cascade pattern is more explicit, but the semantic approach has better accuracy in the hashtag cascade where the cascade pattern is less directly observable. To investigate such a correlation, we resort to the entropy to describe the degree of explicitness in observable patterns across different cascade mechanisms. A higher entropy indicates that a character has more information in a tweet, which means the content in the tweet is more complicated, and the cascade pattern is less directly observable. As we can see in Table 1, the time length of most cascades are transient, n_w (w=2) is only 50% of n_w (w=1) for all these diffusion mechanisms. Thus, to get enough data for training set, we focus on the cascades with 8 weeks and above and conduct the experiment for 1 8 weeks' training period. If a cascade continues to diffuse in the future periods, the desired output label is "True", otherwise, it is "False". We adopt the "Leave-one-out" method to segment the training set and the test set.

Inter-relations

To investigate the inter-relations between the semantic approach and the transactional approach, i.e. whether they are independent or overlapping, and whether it will be more effective if we combine different methods together, we integrate the binary prediction results through the logical conjunction and disjunction, and then compare the prediction accuracy of the integrated prediction result with the individual prediction result. The logical conjunction and disjunction of prediction results are shown in Equation (6).

$$\lambda_{conj} = \lambda_{Smt} \wedge \lambda_{Tran}, \quad \lambda_{disj} = \lambda_{Smt} \vee \lambda_{Tran}, \qquad (6)$$

where λ_{Smt} denotes the output label ("True" or "False") of sematic based approach with regard to whether a cascade continues to diffuse or not , λ_{Tran} denotes the output label of transactional based approach (TF-IDF and DTM).

5 Results and Analysis

In this section, we report the findings based on the methodology described in Section 4 to compare between two transactional approaches (see Section 3.2), i.e. predicting information cascades based on the TF-IDF features and topic (extracted from DTM) features, and one semantic approach (see Section 3.1), i.e. predicting information cascades based on the five semantic features investigated in this paper– sentiment, controversy, content richness, hotness and trend momentum. A detailed comparison of these two approaches w.r.t prediction accuracy and their inter-play is reported.

5.1 Prediction Accuracy Comparison

Figure 1 displays the prediction accuracy of both the transactional and semantic based diffusion methods for each of the three cascade mechanisms: hashtag cascade, url cascade, and retweet cascade. Note that the difference mainly lies in the recall rate, while the precision performance is more or less the same. The precision is about 10%∼20% less than the recall rate at each training period, this indicates that these methods give slightly more false positive predictions (e.g. information that is not diffused in the next time slot but predicted as diffused) than false negative predictions (e.g. information that is diffused in the next time slot but predicted as not diffused). In other words, these methods tend to have a more precise coverage of the truly diffused cascades. In the hashtag cascade scenario, in particular, almost all the information that is diffused in the next time slot has been predicted (i.e. recall near 100%), while undesirably introducing some false positive predictions of the non-diffused cascades. Hence, the following discussions mainly focus on the recall rate.

We can see from Figure 1 that the performance of different methods under distinct diffusion mechanisms vary. For the hashtag cascade, the semantic approach outstandingly outperforms the transactional approach, the recall rate mostly remains above 90% (and the score for the training period equal to 1 week and 2 weeks is 54% and 78%, respectively), while the scores for TF-IDF and DTM are only 22%∼48% for 1-week training period, and climb up to 80% around 6-week training period, and reach 95% at 8-week training period. The precision scores are more or less the same for all the three methods, which are between 20%∼80%. However, in the url cascade and the retweet cascade, the transactional models have notably better performance, in particular, TF-IDF has the best recall rate in both cases, with 60% at 1-week training period for both cascades, and 20%∼95% for longer training periods. DTM, on the other hand, outperforms the semantic approach in the retweet cascade scenario, with 5%∼20% higher at each training period, but has more or less the same performance in the url cascade, with 47%∼59% for DTM, and 46%∼62% for the semantic approach.

This discrepancy in the performance of prediction accuracy is related to the direct observability of the patterns in different cascade mechanisms, see Table 2 for the average information entropy. The results indicate that in the least obvious pattern scenario: the hashtag cascade with the highest entropy 2.29 ∼4.34, the semantic based prediction has the best prediction quality. In the cascades with more directly observable patterns, i.e. the url cascade and the retweet cascade, with the average entropy 2.99∼3.4 and 2.17∼2.56, respectively, TF-IDF has the best prediction accuracy. This means that when the cascade pattern is more directly observable, the feature extraction based on keywords (as in the case of TF-IDF) is more effective in predicting whether the cascade will continue to diffuse in the future, than both the semantic features and the features based on latent topic distributions. The results demonstrate that the performance of the semantic approach and the transactional approach is correlated with the content complexity in different cascade mechanisms.

The complexity analysis in Table 3 shows that the semantic approach is the fastest, TF-IDF in the middle, and DTM

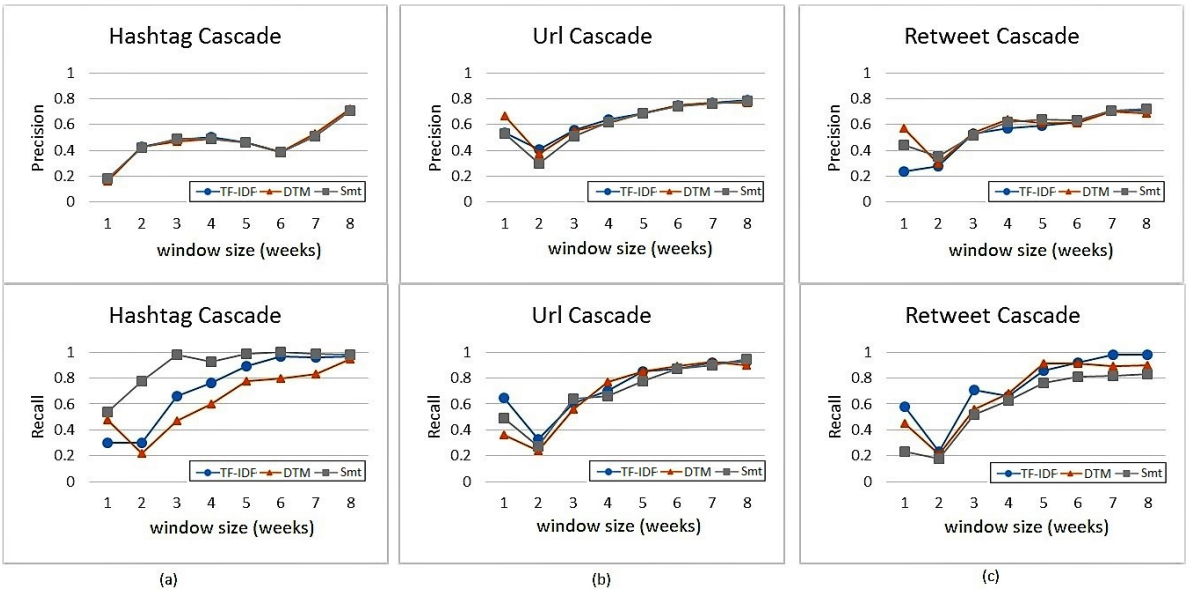

Figure 1: Prediction accuracy analysis of transactional approach and semantic approach.

w (weeks)	Hashtag Cascade	Url Cascade	Retweet Cascade
$w = 1$	2.29	2.99	2.56
$w = 2$	3.05	3.02	2.22
$w = 3$	3.45	3.14	2.17
$w = 4$	3.73	3.24	2.19
$w = 5$	3.93	3.3	2.18
$w = 6$	4.1	3.37	2.22
$w = 7$	4.23	3.4	2.26
$w = 8$	4.34	3.4	2.39

Table 2: Content Entropy w.r.t Different Cascade Mechanisms

	TF-IDF	DTM	Smt
Time Cost (ms)	1920.22	11053.38	32.47

Table 3: Complexity Analysis

costs the longest time. Therefore, the semantic approach is more preferable in the hashtag cascade, and TF-IDF is more preferable in the url cascade and the retweet cascade.

5.2 Inter-relations Analysis

We analyze the inter-relations between the semantic approach and the transactional approach by studying the logical conjunction and disjunction of their prediction results. To enable a finer view in the variation in each week in the 8-weeks training period, here we focus on window size = 1 week, and the results are exhibited in Figure 2.

In Figure 2 (b), we can see that the semantic based approach and the transactional approach do not function independently, but bear a distinctive similarity. This observation holds for both the transactional methods (TF-IDF and LDA) in conducting the intersection analysis. For the hashtag cascade, the intersections of semantic and transactional approaches (Smt-TFIDF and Smt-DTM) lead to only 0.05% less than the smaller recall rate of the original two approaches, with the recall rate above 55% for Smt-TFIDF and above 86% for Smt-DTM. For the url cascade, the intersection methods Smt-TFIDF and Smt-DTM result in 5-20% less than the

	Hashtag Cascade	Url Cascade	Retweet Cascade
Smt-TDIF			
$\frac{\lvert TP(\lambda_{conj})\rvert}{\lvert TP(\lambda_{Smt})\rvert}$	89.42%	74.09%	67.86%
$\frac{TP(\lambda_{conj})\rvert}{\lvert TP(\lambda_{TFIDF})\rvert}$	97.87%	75.70%	49.88%
Smt-DTM			
$\frac{TP(\lambda_{conj})\rvert}{\lvert TP(\lambda_{Smt})\rvert}$	94.68%	69.52%	62.52%
$\frac{TP(\lambda_{conj})\rvert}{\lvert TP(\lambda_{DTM})\rvert}$	97.90%	74.75%	58.84%

Table 4: Average Overlapping Percentage

smaller recall rate of the original, with the recall rate above 13∼81% for Smt-TFIDF, and 9%∼83% for Smt-DTM. For the retweet cascade, the intersection models have 13% less than the smaller recall rate of the original, with the recall rate above 7∼88% for Smt-TFIDF, and 7%∼87% for Smt-DTM.

Further, we note that the shape of the conjunction set curve (see Figure 2 (b)) for the hashtag cascade and the url cascade bears a distinctive similarity to the original TF-IDF curve and the original DTM curve (see Figure 2 (a)), while the shape of the logical conjunction integration curve differs greatly from the original for retweet cascade. This is because the conjunction integration takes up a consistently higher percentage in the transactional approach than in the semantic approach for the hashtag cascade and the url cascade. However, the conjunction integration for the retweet is more similar to the semantic approach at the first half (i.e. till the 4th week), and more similar to the transactional approach at the latter half. The average overlapping degree between the semantic approach and the transactional approach in Table 4 demonstrates this trend as well. Moreover, for the hashtag cascade, the overlapping degree is the highest, above 89.42%. For the url cascade, it is between 69.52% 75.7%, and for the retweet cascade, it is the lowest, only 49.88% 67.86%.

In Figure 2(c), we can see that the recall rate after integrating the semantic approach and the transactional approach

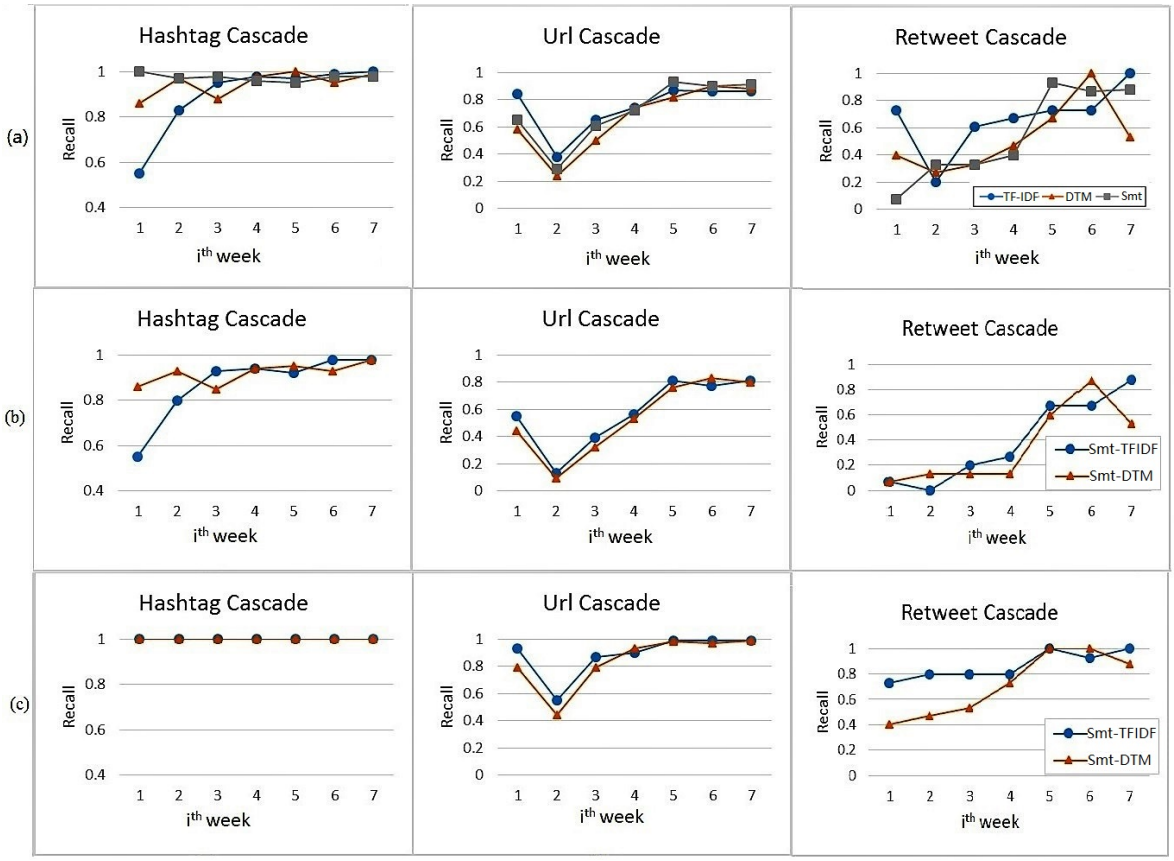

Figure 2: Inter-relation analysis of prediction results of semantic and transactional models: (a) individual, (b) conjunction, (c) disjunction.

leads to a great enhancement. The prediction accuracy for all the three cascades reaches almost 1. For the hashtag cascade, the recall rate achieves 100% for all the training periods. For the url cascade, the recall rate gets above 80% when the 1-week sliding window moves to the 3rd week, compared with 78%, 67%, 76% for TF-IDF, DTM, and Smt, respectively. The recall rate reaches above 90% when the 1-week sliding window moves to the 4th week, compared with 85%, 85%, 84% for TF-IDF, DTM, and Smt, respectively. For the retweet cascade, Smt-TFIDF has better recall rate than Smt-DTM, achieves 80% starting from the 1-week sliding window moving to the 2nd week. The recall rates of both Smt-TFIDF and Smt-DTM reach above 73% when the 1-week sliding window moves to the 4th week, compared with 67%, 47% and 40% for TF-IDF, DTM, and Smt, respectively.

The above analysis shows that the accurate predictive results of the transactional approach and the semantic approach are not independent but overlapping. When integrating both models together, it leads to much more accurate prediction results. In particular, combining Smt and TF-IDF has better performance in terms of both the prediction accuracy and efficiency for all the three ways of information cascades.

6 Conclusion

In this paper, we evaluate the performance difference between the semantic based approach and the transactional based approach in Twitter data w.r.t different diffusion mechanisms, namely, the retweet cascade, the url cascade, and the hashtag cascade. The novelty of this paper lies in (i) showing the performance of the semantic approaches and the transactional approaches in cascade prediction is highly correlated with the content complexity of different cascade mechanisms: the semantic approach has better prediction accuracy in the hashtag cascade with higher entropy, while the transactional approach, in particular, TF-IDF has better prediction accuracy in the url cascade and the retweet cascade with lower entropy; (ii) demonstrating that combing both approaches leads to a large improvement in the prediction accuracy with a detailed analysis of their inter-relations; (iii) showing that integration of transactional approaches and semantic approaches complement each other and bear a significant value for the retweet and the url cascade through the inter-relations analysis; (iv) providing a quantitative analysis of two mainstream focuses—traditional transactional approaches and recently emerged semantic approaches in resolving data heterogeneity and interdependence problem in big data analysis.

In the future, there are at least two directions. First, investigate the integration ways of semantic features into diffusion models to leverage the advantages of both approaches. Second, the semantic features investigated in this paper focus on the diffused cascades, the approach can similarly be applied to describe users to explore the user influence in the cascades.

Acknowledgments

This work was partly supported by Griffith University's 2018 New Researcher Grant, with Dr Can Wang being Chief Investigator. This research was partly supported by BAE Systems.

References

[Akbari *et al.*, 2016] Mohammad Akbari, Xia Hu, Nie Liqiang, and Tat-Seng Chua. From tweets to wellness: Wellness event detection from twitter streams. In *AAAI'16*, pages 87–93, 2016.

[An *et al.*, 2014] Xiaoran An, Auroop R. Ganguly, Yi Fang, Steven B. Scyphers, Ann M. Hunter, and Jennifer G. Dy. Tracking climate change opinions from twitter data. In *KDD'14*, pages 1–9, New York, 2014.

[Bai *et al.*, 2015] Xiang Bai, Song Bai, and Xinggang Wang. Beyond diffusion process: Neighbor set similarity for fast re-ranking. *Information Sciences*, 325:342 – 354, 2015.

[Benhardus and Kalita, 2013] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9:122–139, 2013.

[Blei and Lafferty, 2006] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML'06*, Pittsburgh, Pennsylvania, USA, 2006.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2:1–8, 2011.

[Bradley *et al.*, 1999] Margaret M. Bradley, Peter J. Lang, Margaret M. Bradley, and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *Center for Research in Psychophysiology, University of Florida, Gainesville, Technical Report C-1*, 1999.

[Cheng *et al.*, 2014] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon Kleinberg, and Jure Leskovec. Can cascades be predicted? *WWW'14*, pages 925 – 936, 2014.

[Chew and Eysenbach, 2010] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5:e14118, 2010.

[Guerini *et al.*, 2011] Marco Guerini, Carlo Strapparava, and Gozde Ozbal. Exploring text virality in social networks. In *ICWSM'11*, pages 506–509, Barcelona, Spain, 2011.

[Hoang and Mothe, 2017] Thi Bich Ngoc Hoang and Josiane Mothe. Predicting information diffusion on twitter – analysis of predictive features. *Journal of Computational Science*, pages Available online 28 October 2017, doi = https://doi.org/10.1016/j.jocs.2017.10.010, 2017.

[Hong *et al.*, 2011] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *WWW'11*, Hyderabad, India, 2011.

[Kobayashi and Lambiotte, 2016] Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *ICWSM2016*, pages 191–200, 2016.

[Masucci *et al.*, 2011a] Adolfo Paolo Masucci, Alkiviadis Kalampokis, Victor Martínez Eguíluz, and Emilio Hernández-García. Extracting directed information flow networks: An application to genetics and semantics. *Physical Review E*, 83:026103, 2011.

[Masucci *et al.*, 2011b] Adolfo Paolo Masucci, Alkiviadis Kalampokis, Victor Martínez Eguíluz, and Emilio Hernández-García. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PLoS ONE*, 6:e17333, 2011.

[Naaman *et al.*, 2011] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, pages 902–918, 2011.

[Sprenger *et al.*, 2013] Timm O. Sprenger, Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20:926—-957, 2013.

[Takahashi *et al.*, 2011] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi. Discovering emerging topics in social streams via link anomaly detection. In *ICDM'11*, pages 1230–1235, 2011.

[Taxidou and Fischer, 2014] Io Taxidou and Peter M. Fischer. Online analysis of information diffusion in twitter. In *WWW'14*, pages 1313–1318, Seoul, Korea, 2014.

[Tononi and Sporns, 2003] Giulio Tononi and Olaf Sporns. Measuring information integration. *BMC Neuroscience*, 4:31, 2003.

[Wang and Wang, 2007] Xiaojun Wang and Jianwu Wang. Learning information diffusion process on the web. In *WWW'07*, 2007.

[Wang *et al.*, 2014a] Can Wang, Longbing Cao, Eric Gaussier, and Dan Luo. Coupled behavior representation, modeling, analysis, and reasoning. *IEEE Intelligent Systems*, 29(4):62–80, 2014.

[Wang *et al.*, 2014b] Sen Wang, Xiaojun Chang, Xue Li, Quan Z. Sheng, and Weitong Chen. Multi-task support vector machines for feature selection with shared knowledge discovery. *Signal Processing*, pages 746–753, 2014.

[Wang *et al.*, 2015] Sen Wang, Feiping Nie, Xiaojun Chang, Lina Yao, Xue Li, and Quan Z. Sheng. Unsupervised feature analysis with class margin optimization. *ECML-PKDD*, pages 383–398, 2015.

[Zhao *et al.*, 2015] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. *SIGKDD'15*, pages 1513 – 1522, 2015.

[Zhao *et al.*, 2017] Zhou Zhao, Hanqing Lu, Deng Cai, Xiaofei He, and Yueting Zhuang. Microblog sentiment classication via recurrent random walk network learning. In *IJCAI'17*, pages 3532–3538, 2017.