A string kernel based Information Retrieval approach for tweet-validation through NEWS

Muhammad Rafi[1], Fizza Abid[1], Anum Mirza[1], and Hamza Mustafa Khan[1]

National University Of Computer & Emerging Sciences - FAST, Karachi, Pakistan
muhammad.rafi@nu.edu.pk
http://www.khi.nu.edu.pk

Abstract. Micro-blogging websites like Twitter are very popular among internet users, over 100 million tweets are posted every day. The websites are active in efficient dissemination of information pertinent to any emergency like flood and earthquake. Recent research proved that these platforms can effectively be used for monitoring, evaluations and coordinating relief operations in such situations. One of the very critical issue of such applications is to identifying the validity of these posts automatically during the emergency situation as factual information or rumors. The idea is to verify the tweet from some other authentic news source. Forum for Information Retrieval Evaluation (FIRE 2018) edition included a shared task for Information Retrieval from Microblogs during Disasters (IRMiDis). The subtask 1, is identifying the tweets from their content as fact or fact-checkable tweets. The main idea of this task is to identify the validity of the tweets so that the rumors or baseless situational tweets can be filtered from the context of monitoring and response activity of such challenging emergency situations. The paper proposes a string kernel based information retrieval approach for tweet-validation through NEWS. Our approach is based on two-steps information retrieval. In first step, we consider a given tweet as a query and find a best matching headline using Aho-Corasick algorithm with a score greater than . In the next step, we matched the content of the news with the tweet using cosine similarity, if this score is also greater than , it means that we have a supporting news item for the given tweet. We learn the values of α=0.11 and β=0.25 through experimentation. Our proposed approach performed second best in the competition with overall NDGC 0.667.

Keywords: String kernel approach · Aho-Corasick · Vector space model · tweet validation

1 Introduction

Twitter is considered as a chief communication platform in critical and disaster situations according to the scholars and the practitioners. It has been observed that during hurricanes and real time recovery events, twitter is the most used platform. The significance of twitter cannot be denied as it alerts individuals and help them to recover from such disasters. In crisis situations, through twitter,the

voice of the common people reach to the higher authorities and they will be able to react in best possible manner[6].

However, the tweets can be classified as valid or rumors. Automatic rumor detection is very much dependent on authentic source of information to be met for surety of information. Other challenges include semantic information processing, variational or piecewise information handling, biasness of the information and initiation of the information. At times, rumors are created intentionally to mislead the audience, in such situations, it is complicated to understand the semantics completely. Rumors are of multiple types depending upon the style and language; thus, different algorithms will be required [5]. Otherwise, algorithms trained on limited data will fail due to training biases.

In order to detect valid tweets, the hashtags/keywords will be extracted from the tweets and then they will be compared with the news sources because news sources are authentic and they will easily validate the tweets. NEWS items are the most credible source from which the authenticity and validity of the tweets can be checked. Only if the tweet is based on reality and is considered as a fact, only then it will be discussed in the news. This approach is simple yet effective.

Forum for Information Retrieval Evaluation (FIRE 2018) edition launched a shared task for Information Retrieval from Microblogs during Disasters (IRMiDis). The main approach of subtask 1 is to identify the validity of the tweets so that the rumors or unauthentic situational tweets can be filtered from the context of monitoring and response activity of such challenging emergency situations [1].

2 Literature Review

A variety of communication platforms are evolved with increasing demand of internet enabled communication. Conventional media outlets; for instance, newspapers, radio, and TV for the news, are no more use in the modern era. Nowadays, one of the platform i.e. Twitter has been widely used during a catastrophic situation, such as natural disasters, hurricanes, earthquakes etc .Twitter was proven effective during last many catastrophic events for dissemination of timely information, relief operations monitoring, and response to it. Twitter also possess the tendency to enhance survival during Tornado-related disasters [4].

One of the trivial approach for tweet validation is to see whether this information is available from some authentic news source. Motivated from this thought, we proposed a two-fold string kernel based approach for the problem. The first part of one approach find the similarity of tweet and news using Aho-Corasick algorithm. This algorithm has been utilized for matching keywords/hashtags extracted from tweets with the headlines of the news titles. This string kernel based algorithm is extremely beneficial as its searching phase is straightforward and target each and every occurrence of string [8]. The algorithm creates a finite state machine between various internal nodes. This algorithm is extremely fast and accurate as backtracking is not required [7] [9].

The text is transformed into the vector space model so as to later compute cosine similarity. Vector space model transforms the text (each distinct term) into the vector. For text, the basic idea of a vector space model is to consider each individual term as its own dimension. If we have a document D which has words of length L, then w_{i} is the i^{th} word in D, where I ϵ[1...M] i ϵ[1...M]. Moreover, the group of words in w_{i} are called vocabulary or the term space which is denoted by V. With vector space model (VSM), it is simple to measure similarity between two documents. It is also utilized for document encoding (tf-idf)[3][10].

Later, to compute similarity between the tweet and the news content, the cosine similarity measure has been used. Cosine similarity is the angle between two associated vectors. In information retrieval and related topics, cosine similarity is a commonly used metric. In this metric, text is represented as a vector of terms and the similarity between two texts is obtained from cosine value between two vectors of terms [2][3]. In this metric, text is represented as a vector of terms and the similarity between two texts is obtained from cosine value between two vectors of terms [10].

It was a challenge to learn the accurate value of alpha (similarity between tweet keywords/hashtag and the headlines) and β (cosine similarity between tweet and news content). We used hit and trial method to learn the accurate value of alpha and beta. On α=0.11 & β=0.25, the results were satisfactory.

3 Proposed Approach

The problem is formulated as Information retrieval problem. Given a tweet, we look for the provided news items, to see whether it has some supporting news or not. There are two distinct parts of a news item (i) NEWS Title or heading and (ii) NEWS content. Our approach uses two- fold similarity measures for identifying factual tweets. At first we compute a string based similarity form tweet and NEWS heading using Aho-Corasick algorithm for string-searching, the complexity of this approach is linear as NEWS headings are short text, it is quite quick. All the news items that have a similarity value higher than α (a fixed value we learn from the data), we compute the similarity of NEWS content using simple vector based cosine similarity. If a tweet has a supporting news item(s) and it has a similarity score higher than beta (another fixed parameter). The news item is retrieved as supporting, we arrange all supporting news with decreasing order of similarity as rank scores.

The algorithm is given below. It takes as input T: Set of tweet= t_1 , t_2,...., t_{n} and N: set of NEWS= n_1, n_2,...., n_{k}, where each n_{i} comprise of n_1 : <h_{i} , c_{i}> . The algorithm classify the tweet as factual or non-factual and yield output in the form of t_{k} : tweet and N_{v} : supporting news= n_{v1}, n_{v2},...., n_{vp}.

foreach $t_i \in T$ **do**
  Preprocess(t)
  foreach $n_i \in N$ **do**
    Preprocess(n_i)
    α = StringKernelMatch(t_i, n_i, <h_i, c_i>)
    if α >*threshold* α **then**
      β = cosinesimilarity(t_i, n_i, <h_i, c_i>);
      if β <*threshold* β **then**
        t_k=t_i;
        N_u=N_u;
      end
    end
  end
  return (t_k,N_v);
end

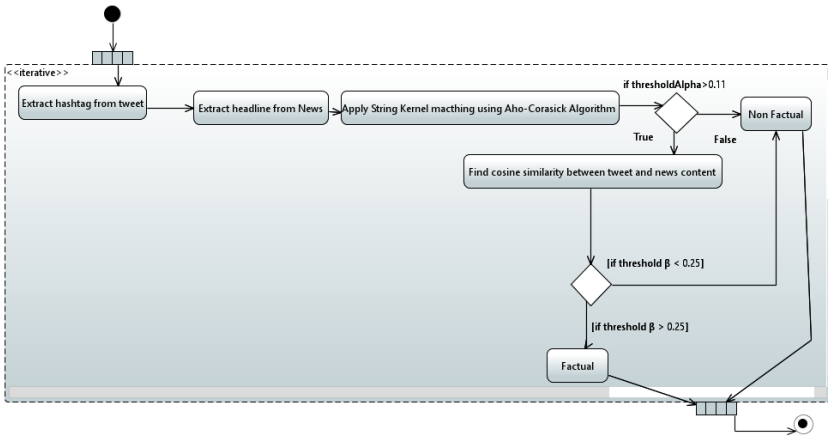

Fig. 1. Flow of the algorithm

Table 1. Example of fact checkable tweets.

Tweet ID	Text	Factuality Score
591974898139987968	RT@Economynext: #SriLanka to fly emergency medical help, food to #earthquake -struck #Nepal #lka #Economynext http://t.co/t6F2RXD4tj http:âĂę	0.5

Table 2. Example of not fact checkable tweets.

Tweet ID	Text	Factuality Score
591981330839130112	RT @dillidikudii: Whoever is in Delhi. Go donate clothes, meds, mineral water etc at the Nepal embassy or Red Cross.	0.00

4 Experimental Studies

The dataset contained 5006 tweets. And, for validation of tweets, news items were also present in the dataset. In order to test our approach, we first run the algorithm on 100 tweets. The first submission is based on tweet similarities with NEWS Headlines and NEWS Content. Aho-Corasick is used for tweet and headline similarity computation (α) and if the value of α is higher than 0.11, the similarity of content with the tweet, is computed using cosine similarity (β). If it is higher than 0.25, we retrieved all such news items as a supporting news items in decreasing order of similarity scores. Whereas, in Run #2, there are hash-tags in the tweets, these has-tags are extracted from tweets and hashtag based similarity is computed with these hashtags and the news headlines. If these values are higher than 0.11, we compute again the similarity of content using cosine similarity if it is higher than 0.25, we retrieved all such news items as a supporting news items in decreasing order of similarity scores. On the other hand, Run #3 combines both the techniques mentioned in Run 1 and Run 2 but the results were not effective.

5 Results & Discussion

The algorithm was initially tested on 100 tweets to automatically and manually test the string kernel based approach. On executing Run 1, the accuracy was 27%. On manually evaluating the results, we observed that the keywords from the tweets are not that accurate as up till yet no library has been designed to extract keywords from short text and research is in process. We have used the library RAKE, which is used for long text. Overall NDCG score was not satisfactory. Run 2 showed relatively better scores. It was based on hashtag approach. The accuracy was 70%. Various tweets did not contain hashtags; otherwise, the accuracy would have been increased. Lastly, the accuracy of Run 3 was 28%; however, the NDCG score was 0.3108, which is not efficient.

Table 3. Evaluation Results

	Precision @ 100	MAP @ 100	NDCG @ 100	NDCG Overall
Run#1	0.2700	0.0068	0.2009	0.3208
Run#2	0.7000	0.0396	0.5723	0.6676
Run#3	0.2800	0.0054	0.1785	0.3108

6 Conclusion and Future Work

A lot of research in the area of fact checking of information has been done during last 10 years. In Future, the idea to extract keywords from the tweets can be used for this approach. One of the trivial approach to information validation is to verify it through authentic news sources. Besides, identifying reliable news items related to tweet and computing a credibility score is an active area of research We have proposed a string kernel based approach for this task. Our proposed system was ranked second as in the competition. There are two possible extension of this approach we foresee 1) Semantic matching of tweet and news content and 2) word sense disambiguation would definitely increase the performance of the system.

References

1. Basu, M., Ghosh, S., Ghosh, K.: Overview of the FIRE 2018 track: InformationRetrieval from Microblogs during Disasters (IRMiDis). In: Proceedings of FIRE 2018 - Forum for Information Retrieval Evaluation (December 2018)
2. Pradhan. N., Gyanchandani, M., Wadhvani,R .: Article: A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications, **120**(9), 29-34 (2015)
3. Price. S, Flach, P.A , Spiegler. .: SubSift: a novel application of the vector space model to support the academic research process. PMLR,(2010)
4. Niles, M. T., Emery, B. F., Reagan, A. J., Dodds, P. S., Danforth, C. M..: Average individuals tweet more often during extreme events: An ideal mechanism for social contagion. arXiv preprint arXiv:1806.07451 (2018)
5. Cao, J., Guo, J., Li, X., Jin, Z., Guo, H. and Li, J.: Automatic Rumor Detection on Microblogs: A Survey. arXiv preprint arXiv:1807.03505 (2018)
6. Takahashi, B., Tandoc Jr, E.C. and Carmichael, C.: Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. Computers in Human Behavior **50** 392–398 (2015)
7. Hasib, S., Motwani, M. and Saxena, A.: Importance of aho-corasick string matching algorithm in real world applications. international journal of computer science and information technologies **4**(3),467–469 (2013)
8. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C.: Text classification using string kernels. Journal of Machine Learning Research **2**(2),419–444(2002)
9. Aho, A.V. and Corasick, M.J.: Efficient string matching: an aid to bibliographic search. Communications of the ACM **18**(6), 333–340 (1975)
10. Paul,M.: Drugs and Popular Culture. Willan, (2013)