Expert Systems With Applications 168 (2021) 114303

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Characteristics of viral messages on Telegram; The world's largest hybrid public and private messenger

Arash Dargahi Nobari, Malikeh Haj Khan Mirzaye Sarraf, Mahmood Neshati *, Farnaz Erfanian Daneshvar

Faculty of Computer Science and Engineering, Shahid Beheshti University:G.C, Iran

A R T I C L E I N F O

Keywords:
Telegram
Instant messaging
Sentiment analysis
Social sensing
Viral message

A B S T R A C T

Telegram is a new Instant Messaging application providing key features for both public and private messaging. Telegram is similar to group broadcast or micro-blogging platforms, while on the other hand, it has features of ordinary Instant Messaging applications such as WhatsApp. In this paper, investigating a real dataset crawled from Telegram, we provide several observations which can explain the information flow, business model of content providers, and social sensing aspects of Telegram. The crawled dataset which is manually labeled by six persons contains two months of public messages of selected Telegram channels. Moreover, we introduce the viral messages in instant messaging services and propose formal definition of these messages as well as deeply analyzing their characteristics and features. Detection of virality characteristics of messages in Telegram can be beneficial for both end-users and digital marketers. Consequently, we propose statistical and word embedding approaches to detect viral messages and their sentiment and message category.Our experiments indicate that the word embedding approach can significantly outperform other baseline models.

1. Introduction

Over the recent years, there has been a growing interest in instant messaging (IM) applications such as Telegram, WhatsApp, and so on. These IMs play a key role in today's world in terms of communication methods that they have provided. With the advent of mobile internet and technologies such as WiFi and 4G, traditional services like SMS (short messaging service) and MMS (multimedia messaging service) have faded away and are about to decline. On the flip side, the use of IM services has increased significantly. These services give users a variety of interesting options. For example, real-time text, file transmission, voice/video call, and forwarding, through which users can send their desired content to their friends list or even groups that they have joined.

In the literature, different studies have been made in the context of IMs and their effects on the ways they have made people connected to each other. For instance, the authors have compared SMS with WhatsApp IM in Church and de Oliveira (2013), the usage pattern of Snapchat Piwek and Joinson (2016) and user behavior analysis in mobile Internet (Yang et al., 2015) as well as users behavior patterns in WeChat and WhatsApp social messaging groups (Qiu et al., 2016; Seufert et al., 2015) have also been studied. Moreover, the usage of IMs has not been restricted to users, but also other aspects of them including IoT (de Oliveira et al., 2016) and healthcare (Ghaffari et al., 2017) have been investigated. It is worth mentioning that a large portion of these studies analyzed the problem using surveys and questionnaires which were distributed among a limited number of people.

More recently, network structural aspects of Telegram instant messaging service has been analyzed in Dargahi Nobari et al. (2017a), where the authors have utilized real data, crawled from Telegram network, in their analysis. Their investigations indicate that the characteristics of the network in Telegram are different from other social media. Furthermore, common metrics of quality evaluation, like PageRank, cannot be applied to this network.

Today, some of these networks have gone beyond the concepts of simple instant-messaging. In addition to the exchange of real-time text, voice/video, and file transmission, they have provided their users with some facilities by which users can share messages with large audiences. This facility is known as *channel*. Channels are platforms that have been introduced by some IM networks such as Telegram, and Line Messenger. The goal behind channel platforms is to provide users with an option to broadcast their public messages to large audiences. Channels are also used for official broadcast from organizations and corporations (Jalilvand & Neshati, 2020). Furthermore, one of the key actions of IM users is the option of sharing messages. This can be

* Corresponding author.
E-mail addresses: a.dargahinobari@mail.sbu.ac.ir (A. Dargahi Nobari), m.sarraf@mail.sbu.ac.ir (M.H.K.M. Sarraf), m_neshati@sbu.ac.ir (M. Neshati), f.erfaniandaneshvar@mail.sbu.ac.ir (F. Erfanian Daneshvar).

https://doi.org/10.1016/j.eswa.2020.114303
Received 20 March 2020; Received in revised form 15 October 2020; Accepted 9 November 2020
Available online 12 November 2020

performed through forwarding messages from channels to the friends list, or private groups of users.

At first glance, the concept of channel seems to be highly similar to famous micro-blogging platforms such as Twitter. However, the distinction point is the difference between user actions in these platforms, which we aim to discuss thoroughly in this paper.

Our experiments show that messages of a channel are not all of the same importance for users and some of them are forwarded to other users more frequently. This category of messages is often referred to as “viral messages”. In other words, with the introduction of “channel” concept in IMs, public messages circulate in group/private chats. This new information flow has some special features which do not exist in other networks such as Twitter.

1.1. Research problem

In this paper, we investigate several problems related to viral message on Telegram network. Specifically, in our first research question we focus on (RQ1) Detection of viral messages and their category and sentiment. The second research question investigates (RQ2) The role of textual features on message classification. Next, we answer the question (RQ3) Is there any specific relation between sentiment and virality of messages? And finally, (RQ4) What is the effect of a channel’s activity and popularity on the virality of its messages?

The basic idea of our work has been recently published as a short paper in CIKM conference (Dargahi Nobari et al., 2017a). By crawling a new dataset with more messages containing detailed information, in this paper, we investigate new and untouched aspects of Telegram.

1.2. Main findings and contributions

The major contributions of this paper are as follows:

- We compare and analyze similarities and differences of Telegram with other social networking and IM apps.
- The information flow of messages, which are published in Telegram channels, is discussed and analyzed.
- The business model of Telegram’s content providers (i.e. channels), as well as its effect on content production, publishing, and further generation of viral messages have been analyzed.
- The characteristics of viral messages from the sentimental aspect have been discussed.
- Finally, a method will be proposed to detect viral messages right after their publishing time. In order to decipher the sentiments of messages, as well as their type, appropriate classifiers, which increase the precision of baseline models, will be introduced.

In this paper, we have collected a new and complete dataset by crawling over 450k public messages on Telegram in the period of 2 months. Of these messages, 18,566 messages were shown to have gone viral. For all of the viral messages, we asked human annotators to determine the type and sentiment of the message. We believe that this process has made our findings interesting and our dataset a very valuable source that has been made publicly available for the use of other researchers.

As we said, the collected dataset contains the messages published on Telegram channels. All channels have public content and their subscribers[1] can access the messages without any restrictions. Thus, our dataset is not private and can be published.

2. Related work

Over the recent years, there has been a growing interest in analyzing social media among researchers. A great deal of work has done on different aspects of Twitter network (Anagnostopoulos et al., 2018; Chatzakou et al., 2017; Gupta et al., 2018; Keib et al., 2018; Kim & Shim, 2014). For instance, Cha et al. (2012) analyzed the influence of Twitter users by employing three measures that capture different perspectives: in degree, re-tweet, and mentions. Moreover, in Alshammari et al. (2020) and Jiménez-Bravo et al. (2019) authors have proposed new user modelings by utilizing relationships among users to provide more accurate recommendation systems. In addition to Twitter, Facebook has been investigated in recent studies. For example, authors have studied users’ behaviors regarding ideologically discordant content Bakshy et al. (2015).

Studying IM networks is a new line in research and it has attracted widespread interest in recent years. Specifically, in Church and de Oliveira (2013), authors have compared users’ behaviors in IM platforms with those in SMS. Research concerning Telegram itself has mostly focused on security aspects (Sušánka & Kokeš, 2017), social influence (Dargahi Nobari et al., 2017a; Negreira-Rey et al., 2017) and forensic analysis (Anglano et al., 2017; Gregorio et al., 2017; Yusoff et al., 2017). In Dargahi Nobari et al. (2017a), authors have analyzed public posts in Telegram network. They crawled a fair volume of public data of Telegram which were divided into two groups including advertisement/non-advertisement. Then, they proposed a method to detect advertisements. Finally, they concluded that Page Rank is not a good measure for detecting valuable nodes in Telegram since there are many advertisement edges in the network graph.

There has been some work on detecting viral messages in social media in previous papers. In Camarero and José (2011), chance of a re-tweeting for a tweet is estimated and the problem is solved with a classification method. Moreover, Cha et al. (2012) have investigated Twitter and defined 3 user groups including sources mass media, ordinary users, opinion leaders. They then expanded on users’ role in information flow and how they make news viral. Furthermore, different factors of success in viral marketing have been explored in literature (Camarero & José, 2011).

A subtle line of related works is concerned with event detection in social media (Almerekhi et al., 2016; Doulamis et al., 2016; Hasan et al., 2018), as well as the concepts related to social sensing. In this kind of papers, the goal is to analyze humans’ behavior as a sensor for transmitting data ("Xiaomei et al., 2018; Yuan et al., 2019). Specifically, in Sakaki et al. (2010), the authors aimed to detect tweets related to a specific event by using the sense that is associated with each tweet. They utilized Kalman filtering or particle filtering and extracting associated features to detect an event (earthquake). Additionally, in Li et al. (2017), by using semantical analysis of words, a real-time method has been proposed to detect novel events in Twitter. Events have been detected with regard to their semantics. Shifting/changing the sense associated with a tweet may make that event important. That is, users can alter the sense of a tweet about a specific topic and make it attractive. To this end, BNgram and sentiment spikes methods have been used. In Hua et al. (2013), authors proposed a semi-supervised system that helps users to automatically detect and interactively visualize events of a targeted type from Twitter. More recently, by exploiting an incremental clustering approach, Hasan et al. (2019) have introduced a new real-time event detection framework on twitter data stream. Besides, some studies have been done to track and detect events from heterogeneous news streams (Mele et al., 2019).

Sentiment analysis in social media has also gained a lot of interest over recent years (García-Pablos et al., 2018; Schouten et al., 2018; Zainuddin et al., 2018). It has widely done on data of Twitter (Severyn & Moschitti, 2015). For instance, in Giachanou et al. (2016), authors focused on tracking sentiment towards different entities, detecting sentiment spikes and on the problem of extracting and ranking the causes of a sentiment spike. Their approach combines LDA topic model with Relative Entropy. Thelwall et al. (2010) proposed SentiStrength, an algorithm to extract sentiment strength from informal English text. In Xia et al. (2015), a learning method based on pseudo-antonym dictionary for sentiment analysis of texts has been proposed. Lately, de Lira et al. (2019) have proposed a method to infer the users’ attendance of large events.

[1] All users have the option to subscribe to channels using their address.

Table 1
Different actions available in Telegram, Twitter and WhatsApp.

Telegram	Twitter	WhatsApp	Comment
Send message to user	Direct message	Send message to user	Direct message is a minor feature in Twitter and not a main functionality.
Send message to group	–	Send message to group	
Send message to channel	Tweet	–	
Read public message	Read tweets	–	Reading messages published in Telegram channels causes view counter to increase.
User message forwarded to user/group	–	Forward	The forwarded message in Telegram includes the author's signature. However, forwarding action in WhatsApp is practically similar to copy–paste action.
Channel message forwarded to user/group	Direct message	–	It is only the case for users in Telegram and there is nothing called "group" in Twitter. In Telegram, when a message is forwarded to groups/channels, the view counter associated with the forwarded message will grow.
Channel message forwarded to another channel	Re-tweet	–	In Twitter, re-tweeting often happens for confirmation of a message. However, according to the analysis carried out in Dargahi Nobari et al. (2017a), this action occurs for commercial purposes in Telegram.
User message forwarded to channel	–	–	In this case, the author's signature is published in channel. It is widely used when the administrator wants to quote from a famous person.
Message reply (private and group chats)	Message reply (direct and group chats)	Message reply (private and group chats)	
–	Comment	–	Users are not allowed to give feedback on published content in Telegram.
–	Like	–	
Join group	–	Join group	Twitter does not support an option to create groups with certain number of users.
Subscribe/leave a channel	Follow/unfollow an account	–	

3. Telegram is different from Twitter and WhatsApp

Telegram is an IM service that, in addition to the private messaging, provides numerous facilities to its users. In Telegram, users can send messages to each other. Encrypted messaging, data storage, media streaming, and so on, all together have made Telegram distinct from similar applications. To better understand this distinction, in this section, after getting acquainted with Telegram, we compare it with Twitter and WhatsApp, and finally present information flow of public content on it.

3.1. Introduction of Telegram and user actions on the network

Communications on Telegram network are not limited to two-way communications (Dargahi Nobari et al., 2017a). The other communications are as follow:

1. **Groups:** Groups allow users to talk to each other collectively.
2. **Bots:** Telegram has offered an open bot API by which developers can create different bots and provides users with appealing options (Telegram, 2020).
3. **Channels:** Telegram Channel is a feature to let administrators broadcast public messages to large audiences. The users can share these contents with their friends list, i.e., possibly other users in the network.

3.2. Comparison of Telegram with other social media

In this section, in order to enlighten different aspects of Telegram, we aim to compare differences and similarities between Telegram, WhatsApp,[2] and Twitter micro-blogging service, which is mostly used as a platform for publishing public content and rarely used as an IM. Our experiments show that Telegram has successfully juxtaposed these sets of features and stands as a new kind of social media between them. In what follows, we compare Telegram with WhatsApp and Twitter in terms of *Users' Actions*, *Content Privacy*, and *Business Model of Content Providers*.

3.2.1. Users' actions

In each of the mentioned networks, end users are able to do a predefined set of actions. In WhatsApp, Users' actions are mostly done on private messages and include sending messages, replying, and so on. On the other hand, in Twitter, most of the users' actions are performed on the public content. For example, users can *tweet* and give feedback (commenting and liking) on others' contents. Our studies reveal that Telegram users' actions can be carried out on both public and private content. Table 1 demonstrates some actions of users in the mentioned networks and further compares them with each other.

3.2.2. Privacy level

As mentioned earlier, WhatsApp, as an IM application, provides users with required features to publish private messages. The importance of private content in Whatsapp is crucial such that author's name is not reflected in a forwarded message. On the flip side, Twitter is known as a public content platform therefore author's name is displayed when he/she *re-tweets*.

Fig. 1 illustrates different aspects of instant messaging in Telegram. Clearly, Telegram provides a combination of public and private messaging. More precisely, the inner circles demonstrate private contents

[2] An IM application which only supports 2-way and group communications.

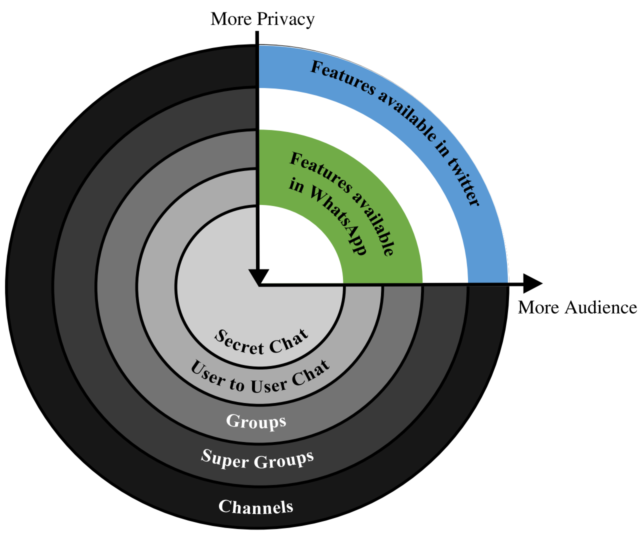

Fig. 1. Comparison of Telegram, Twitter and WhatsApp.

(similar to WhatsApp) while outer circles show public contents (similar to Twitter) in Telegram. From this point of view, Telegram can be recognized as a tool for publishing both public and private contents.

3.2.3. Business model of content providers

As pointed out in the previous section, Twitter and Telegram can be used to publish public content. A key point is the differences in delivering values to content providers in these platforms. In Twitter, content providers produce different contents to deliver them widely among users. In other words, this platform brings lots of audiences and also a media for content sharing. On the other hand, since Telegram has a limited number of users, compared to Twitter, it has made a different value for its content providers. To put it another way, in Telegram's network, public content providers, i.e., channel administrators, are able to monetize by publishing advertisements in their channels. From this viewpoint, every Telegram channel is like a Twitter feed that can expose advertisements to users. Telegram has also a special feature, called view counter, by which channel administrators can figure out how many users have seen their published messages. Channel administrators strive to deliver better content therefore the sharing rate of their messages would increase and accordingly, the view counter of their messages would rise. Precisely, a cycle including *better content* → *more subscribers* → and consequently *more expensive advertisements* describe the revenue model of Telegram channels. In other words, there is a major difference in the concept of content feed between Telegram and Twitter. Twitter is mainly a single-feed platform. Thus, the revenue of advertisement is only for the platform itself, and content providers do not have any share in the profits. However, in Telegram, every channel has a distinct feed which makes it a multi-feed platform.

On the other side, Twitter only supports short text, with an image or a video, whereas Telegram is a rich platform for sharing different types of contents including images, videos, a wide variety of files, location, and so on.

3.3. Information flow of public content on Telegram

As discussed previously, Telegram is quite different from other social media and Microblogging services in terms of types of content, users' actions, and business model of content providers. In fact, utilizing features of both systems, Telegram has introduced a new and distinct information flow.

Fig. 2 demonstrates information flow of Telegram channels. The channel administrators sense and receive events from different sources such as physical environment, social media platforms, and so on. Then, they broadcast the sensed contents in their channels. Channel's subscribers can forward contents to their friends list or groups in which they are member. The contents would virally be forwarded to other users, groups, or channels as well. Forwarding channel's messages leads to an increase in the view counter of the message.

4. Viral message on Telegram

As noted earlier, the combination of public and private platform's features in Telegram has caused the public content of Telegram channels to be published in more private sections such as groups or private chats. In this way, users can receive appealing contents with more ease as human beings are more sensitive to the private/group chats rather than public channels. Keeping this fact in mind, a large portion of messages, that are published in public feeds, i.e., channels or Twitter feed, are overlooked by users. However, those messages being sent from friends through private/group chats, are more probably to be seen and read by the user. Thus, from this point of view, messages in Telegram can be divided into two groups as follows:

- **Viral Messages:** which are sent by audiences of channels to other users or different groups.
- **Ordinary Messages:** which are only read by the users in channels, not sent to others.

In this section, we first provide a formal definition of the viral message. Then, we look at some important observations on Telegram public content. Finally, we define research problems related to viral message in Section 4.3.

4.1. Viral message definition

Fig. 3 shows messages of a channel along with their view counter. The horizontal line signifies the ID number of messages of a given channel in chronological order, and the vertical line shows the view counter of each message 24 h after publishing time. As indicated in the figure, some messages (specified in circles) have quite a peculiar view counter in their neighborhood. In other words, these are those posts whose view counter is a local maximum in a close vicinity.

As reported in Dargahi Nobari et al. (2017a), users' activities in Telegram differ at different times of a day and as a result the view counter should be compared in a close vicinity, not entire data. In Fig. 3, the view counter of message E is meaningfully less than that of messages B and C. However, message E can also be regarded as a viral message because its view count is significantly higher than its neighbor messages.

Fig. 4 indicates the view counter of a sample message in different time slots, after its publishing time. Right after the publishing time, the slope of the view counter diagram is significantly high, showing that most users see the messages immediately after they are published. This occurs since when a message is published, a push notification (PN) is sent to users and the probability of opening these PNs decreases gradually. This figure indicates that approximately 24 h after the publishing time, view counter approaches a fixed number (i.e. the steady state). When this fixed number is higher than the other numbers in a close vicinity, it means that users have acted as a media and forwarded this kind of messages to others.

With regard to the given explanations, the formal definition of viral messages is as follows:

Definition 1 (*Viral Message*)**.** A published message in a Telegram channel is called *viral* iff the associated view counter in steady state is $\alpha\%$ higher than the average of other posts in a vicinity of n posts.

α and n are the parameters of definition, and will be estimated in 7.1.

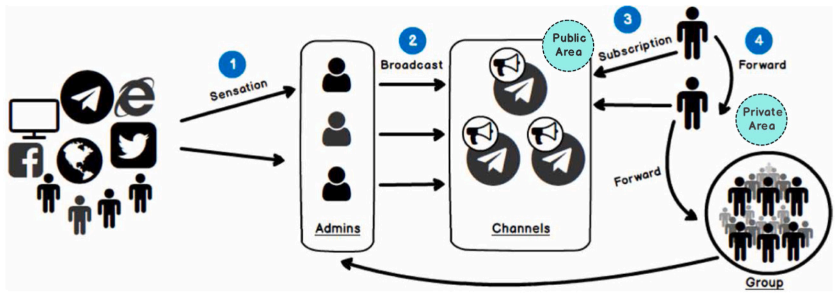

Fig. 2. Information flow of channels in Telegram.

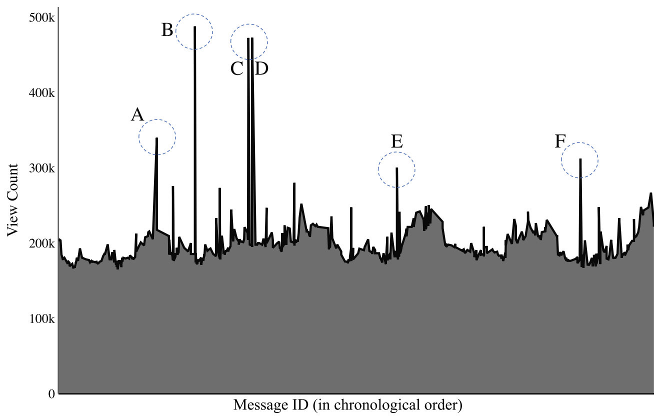

Fig. 3. View Counter of messages in a sample Telegram channel. The data may seem continuous due to the compactness of data, however, they are discrete.

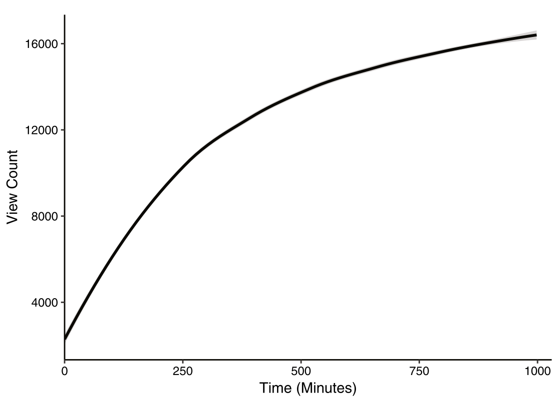

Fig. 4. Growth path of ViewCount for a sample message.

4.2. Observations on viral message

The exclusive features of Telegram in terms of sending and receiving messages in public and private contexts lead to its special properties, some of which we will mention in this section. In the first part, the reason why the viral messages have high impression is described. In the second part, the advertising mechanism in the Telegram channels is explained and in the last part, Telegram functionality is described as a social sensing tool.

4.2.1. Abundance of push notifications and users' reactions

As mentioned before, each channel in Telegram functions like a media which publicizes different information to users. Each message, which has been published on these channels, is sent in the form of Push Notification (PN) to the users. Depending on the type of a channel, the number of PNs, which are sent to users, differs in a specific day, ranging from 0 to over 100 PNs sent daily. If the number of notifications goes higher, users will often ignore them and accordingly, this will decrease the users' sensitivity to the PNs (Sahami Shirazi et al., 2014). In this regard, the user may have hundreds of unread PNs daily. It is worth mentioning that to pass over these PNs, Telegram has provided its users with an option by which users can skip received PNs. Our observations offer that the average view counter (impression) of the messages of a channel is about three times lower than the number of its subscribers. Specifically, we experimented on 20 channels of our test collection.[3] randomly. The average view counters of messages of these channels, 24 h after publishing time, was averagely 38% of their subscribers. This observation asserts that sending messages merely in the sense of push notification is not an appropriate tool for notifying users. Since with the increasing number of PNs, users' sensitivity will be reduced. However, as mentioned in Section 3.3, an option which lets users forward public messages to groups/private chats will increase the chance of message to be seen. With this in mind, the view counter of a viral message of a channel can be several times larger than the number of subscribers of that channel.

4.2.2. Forwarding mechanism and its correlation with the business model of content providers

Fig. 5 shows the distribution of viral messages in terms of topical categories The category of each message is manually determined by human annotators. The process of annotation is described in Section 5.2. The published content in a channel can be either originally posted by the channel administrator(s) or forwarded from other channels. The number of forwarded and original messages are shown in gray and black colors, respectively. The main observations in this figure include: First, in most general categories, the share of original messages exceeds that of forwarded messages. For example, in "event" category, over 90% of messages have been generated by publisher channel. Second, in "advertisement" category, despite other categories, the share of forwarded messages surpasses original messages. Considering the business model of Telegram channels, this can be justified. In this model, channel administrator(s) try to attract more subscribers. Since users can subscribe to every channel free of charge, popular channels start advertising different products to their subscribers for revenue purposes. More accurately, there is a cycle as follows: better content brings in more subscribers and as a result, the channel administrator will demand more fee for advertising purposes. This process describes the revenue model of Telegram channels. Generally, these advertisement messages are once generated and then forwarded to other channels.

4.2.3. Social sensing on Telegram

As discussed earlier, Telegram users, who contribute to forwarding rate of viral messages, act like a media. Clearly, with the detection of unexpected events and catastrophes, users begin announcing others and forward viral messages. Fig. 6 shows the number of viral messages in

[3] The test collection is described in Section 5.

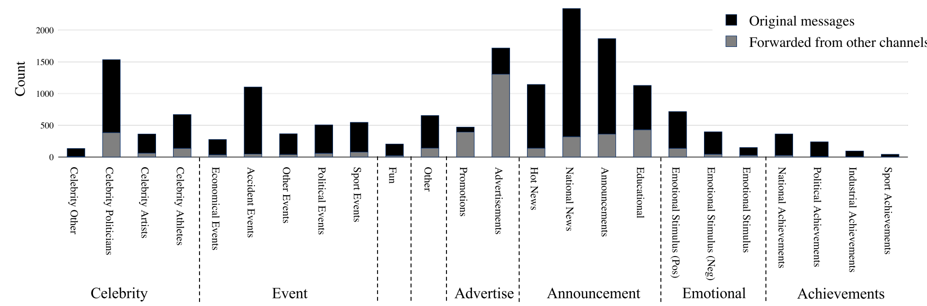

Fig. 5. Comparison of original and forwarded viral messages in each category.

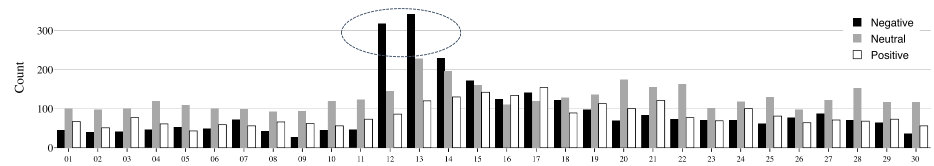

Fig. 6. Sentiments of viral messages sent in Telegram network in the entire days of November 2017.

November 2017 through the channels of our dataset. These messages are shown in black, gray, and white colors according to their sentiment. The sentiment of each message is manually determined by human annotators. The process of annotation is described in Section 5.2. Each color conveys a sense of negativity, neutral, and positivity respectively.

Naturally, neutral messages outnumber both negative and positive messages (daily statistics). Having said that, the number of negative messages (dashed-line in the figure) surpasses the others in particular dates. Our experiments prove that the marked dates are related to a strong earthquake that happened in that period.

Considering observations that have been made in this section, viral messages are of high importance for users in Telegram network. Since each user can be different in terms of their interests, they may be attracted to receive specific category of viral messages. Thus, predicting viral messages right after their publish time is an imperative problem which is discussed thoroughly in the following section.

4.3. Research problems

According to the definition provided in Section 4.1 and observations described in Section 4.2, several research problems have been introduced in this section. These problems are important because they can help us to better understand the message types, i.e., viral and non-viral, their important features, and message flow properties in Telegram.

(RQ1): Detection of Viral Messages and Their Category and Sentiment

In this problem, the question is how can we detect a viral message right after its publishment. In addition to detection of viral messages, it is important to detect their topical category and sentiment.

This is a classification problem in which the input is a raw message and the output is the type of input message and its category and sentiment. What is important in this problem is that merely features that are available right after the message publishment can be used to detect the virality, category and sentiment of messages.

As illustrated in Fig. 5, viral messages cover different categories in Telegram. By identifying the category of viral messages, these messages can be personalized to the user's exact interests and it can improve the user experience of end users.

As illustrated in Fig. 6, an abundance of viral messages with negative sense can account for a recent catastrophe in society. Past studies have shown that recognizing sentiment is important in event detection problem (Paltoglou, 2016). Consequently, understanding the sentiment associated with each message is an important problem, by which other problem like event detection can be solved.

(RQ2): What Role does Textual Evidence Play in Message Classification?

Intuitively, words are expected to play a key role in identification of messages in terms of their virality, sentiment, and category. Having that in mind, our approaches are extracting the most discriminative words for each class of messages and using them as an important feature for the classification task. The effect of textual evidence in the tasks as well as presenting most discriminative words for some sample categories are studied in Section 7.2.1.

(RQ3): Is There Any Specific Relation between Sentiment and Virality of Messages?

In this question, we are looking forward to finding out whether there is any correlation between messages sentiment and virality. To answer this question, we will statistically analyze viral and non-viral of a sample channel to find out any potential relation between sentiment and virality of messages.

(RQ4): What Is the Effect of a Channel's Activity and Popularity on the Virality of Its Messages?

Another interesting question is whether any relation between a channel's popularity, i.e., number of its subscribers, and activity, i.e., the average number of messages published per day, can be observed or not. We have devised two experiments regarding channels' activity and population to answer this question. Our observations and analysis will be presented in Section 7.2.3.

5. Data collection and annotation procedures

With regard to the definition given in Section 4.1, in this section, we aim to describe the data collection procedure of Telegram's selected channels, detection of viral messages, and annotating them with topical and sentiment labels. To this end, we first explain the data collection procedures and then the annotation procedures.

5.1. Data collection procedures

In Telegram, there are many channels with different topics, such as news, education, sports, and so on. In this paper, a special crawler is designed to collect Telegram messages. This crawler is able to fetch messages from the channels we have given as input. To determine the seed channels, we selected 20 people who do not have similar interests and asked them to determine a number of channels that they think are more important/useful.

After reviewing these channels, we finally selected 145 public channels and gave them to the crawler. We crawled these channels over the period of 10-7-2017 to 12-7-2017 (roughly 2 months). These collected channels cover a variety of topics including news, education, sports, and so on. Thus, the crawled messages do not belong to a specific topic.

According to Fig. 4 and what is described in Section 4.1 in the definition of viral message, we selected the viral messages from all crawled ones. The α (difference between a view counter of a message and view counter of other messages in the neighborhood) and n (length of the neighborhood window) parameters in the above definition are set to 10% and 1, respectively. View counters have been calculated 24 h after publish time. With regard to our experiments, these sets of parameters are the best option to detect viral messages.

With these parameters, the number of viral messages was 18,566. This number of messages has been identified out of 451,735 crawled messages.

5.2. Annotation procedures

After generating the dataset, in order to perform experiments, we need to determine the topical category and the sentiment of each message. For this purpose, we used 6 annotators.[4] First of all, each person was asked to independently review the messages and suggest the categories and subcategories related to them. Then, in a joint session with the presence of all annotators, the categories and sub-categories were selected in such a way to maximally cover the topic of all crawled messages. Another criterion to select categories is that the selected categories should minimally overlap with each other.

In joint session, annotators compromised on eight categories which are shown in Table 2. In addition to general categories, different sub-categories have been derived from initial categories to precisely specify different ranges of the category. For instance, *Celebrity* category has been divided into four different categories including *athletes*, *artists*, *politicians*, and *other*.

After determining the categories and sub-categories, it is time to determine the category of each message. To this end, two annotators working separately, were asked to annotate the category associated with each message. Then, we measured the agreement between them using the kappa coefficient (Sim & Wright, 2005). The kappa value was 0.49 which is presumably fine. When annotators reached a full agreement on the category, the message then was annotated. Otherwise, a third annotator, separately and independently, decided on the category of the message. So, the final label of each message has been determined using major voting. We use the same annotators to decide the sentiment of each message as positive, negative or natural. In labeling the sentiment of each message, we use major voting approach described earlier. The annotators were asked to determine the sense of the messages according to the general meaning and the feeling they get by reading them. For example, messages related to accidents, death, earthquake have negative feeling, and messages related to winning competitions, holidays, celebrations, pleasant events have positive feeling. Among 18,566 viral messages, we ended up with 5388, 5125, and 8053 messages as negative, positive, and neutral, respectively.

It should be noted that before performing both annotation tasks, briefing sessions were held with the presence of all annotators, in order to clarify how to annotate a message, a sample set of messages, apart from the collected data set, were annotated collectively; thus people were trained and almost unanimous.

The list of categories with detailed information, as well as all crawled messages including annotated sentiment and categories are publicly available.[5] Table 3 contains a sample of crawled messages with the determined sentiment and category for each one.[6]

Table 2
Categories with related statistics in dataset.

Categories	Description	#sub-cat	#Messages
Announcement	Public announcements issued by channels. For example, a blocked road on a specific date.	4	6844
Events	Natural and unexpected events such as a devastating earthquake	5	2895
Celebrity	Posts that are published by famous celebrities, or literally their opinion on different matters.	4	2824
Promotion	Discounts and special conditions of sale for different productions and services.	2	2733
Emotional	Messages that stimulate public emotion such as pictures of child labors and etc.	3	1382
Other	Anything else	1	833
Achievements	Different achievements in the contexts of sports, nation, and industry. For instance, introduction of a new device by a modern company.	4	768
Fun	Fun and entertaining content like variety of jokes.	1	287
Sum		24	18,566

Table 3
Examples of annotated viral messages. Original messages have been translated from Persian to English.

Message body	Sentiment	Sub-category
Research has shown that if the food plate is blue, the amount of food eaten will be reduced. If you want to lose weight, use blue plates.	0	Educational
Omid Ebrahimi's penalty, which was not declared a goal by the referee's strange mistake, has caused a wave of arguments.	−1	Sport Events
The second person to survive jumping from Niagara Falls has died from slipping on an orange peel.	−1	Hot News
The most beautiful sleeping place for animals belongs to the swans. The swans sleep in the mother's feathers, which is like a bed.	1	Emotional Stimulus

6. Proposed methods for viral message detection - RQ1

As mentioned before, the problems described in our first research question (i.e. RQ1) are basically the classification problems. In this section, we propose statistical and word embedding approaches to solve these problems.

6.1. Statistical approaches

In this family of approaches, for every model, a classifier is created. The input of this classifier is a feature vector which is extracted from each message of the Telegram. The extracted features of every message can be partitioned into textual, and non-textual features. The textual features signify the existence or non-existence of important words in recognizing the label of the input. Non-textual features are

[4] The annotators were students in the information retrieval course who voluntarily participated in the annotation process.

[5] https://tiny.cc/tgds

[6] The selected samples are translated from Persian to English.

Table 4
Non-textual features.

Feature	Description	Type
f_1	Number of words in the message	Numeric
f_2	Media (Image or video) availability	Boolean
f_3	Is a forwarded message	Boolean
f_4	The hour in which the message is sent	Nominal

further shown in Table 4. These features are fixed for all kinds of the mentioned problems, i.e., virality, topical category and sentiment detection, whereas textual features should be extracted separately for each problem. After generating the feature vector for each problem, two famous classifiers (e.g. SVM and Logistic Regression) are utilized for recognizing the label of each message. The feature selection for each of three problems can be solved with two approaches which are Expectation Maximization (Section 6.1.1) and Mutual Information (Section 6.1.2).

6.1.1. Expectation maximization approach

Expectation Maximization (EM) is a common approach to extracting textual features in the classification problem. The main idea behind is that the language model used in the texts of class c (e.g., viral messages in the detection of viral messages problem) differs from the language model used in the documents of other classes (i.e. non-viral messages). It is worth mentioning that those words that are known to be proper *discriminators* are identified throughout an iterative process and then used as textual features. E and M steps of the algorithm are shown in Eqs. (1), and (2) respectively.

E-step: (1)

$$e_w = tf(w,c).idf(w).\frac{\lambda P(w|c)}{(1-\lambda)P(w|c') + \lambda P(w|c)}$$

M-step: $$P_{EM}(w|c) = \frac{e_w}{\sum_w e_w} \quad (2)$$

where c is the class that we aim to differentiate its words from other categories (i.e. c'), and λ is the parameter of algorithm. After running EM and its convergence, words are sorted according to the probability of $p(w|c)$ to extract top-k important words, which are used for creating feature vector of messages. The mentioned algorithm has been run separately for each of three problems to extract proper textual features for each problem.

6.1.2. Mutual information approach

Mutual Information (MI) is a measurement that is defined as how much information the absence or presence of a term leads to making a correct classification decision. MI has been used for many retrieval purposes (Dargahi Nobari et al., 2017b). For each pair of word w and class c, MI is calculated using Eq. (3).

$$MI(c,w) = \sum_{M_c=0,1}\sum_{M_w=0,1} P(M_c, M_w)\log\frac{P(M_c,M_w)}{P(M_c)P(M_w)} \quad (3)$$

where M_c, and M_w are binary variables signifying the occurrence event of class c and word w, respectively. For each class c, the relevance probability of word w to class c is yielded as follows:

$$P_{MI}(w|c) = \frac{MI(c,w)}{\sum_{w'} MI(c,w')} \quad (4)$$

This probability (i.e. $P_{MI}(w|c)$) yields the relevance probability of word w to class c. Supposedly, $P_{MI}(w|c)$ will increase if word w and class c appear together.

In this method, after estimating $P_{MI}(w|c)$ for all words, words are sorted in descending order and top-k words of the list are used to generate feature vector. It should be mentioned that similar to the previous approach, this approaches should be run separately for each of three mentioned problems to identify proper features for each problem.

6.2. Word embedding approach

Recently, the word embedding methods has been used in text classification problems. In some scenarios (Dargahi Nobari et al., 2017b), it outperforms the classical feature extraction and classification models. In this section, we propose a word embedding approach to solve the classification problem mentioned in Section 4.3 (i.e. RQ1). In this approach, each word is embedded in a semantic space with a low number of dimensions. In this space, we expect that semantically similar word have similar representing vectors.

The process of word embedding approach begins with applying topic modeling (i.e. LDA[7]) to the given set of messages which would result in getting a low-dimensional representation of single words. Then, we implement a mapping function from the topic space (i.e. LDA topics) to the class space (for example in our first classification problem, the class labels are viral or non-viral). For notational convenience, we write $P(c|.)$ for the probability distribution over classes. The probability of relevancy of class c, given the word w, is calculated using Eq. (5):

$$P_{WE}(\boldsymbol{c}|w) = \frac{1}{z}e^{W_{LDA}}.W_c + b \quad (5)$$

where W_{LDA} is a $1\times T$ vector showing word w in the topic space (T denotes the number of topics in LDA), W_c is a $T\times x$ matrix which maps topic space representation of word w to the class space,[8] and b is a $1\times x$ vector denoting the prior probability of relevancy of class c to a given word w, and z represents the normalization factor which is calculated as follows: $z = \sum_{j=1}^{|c|}[e^{W_{LDA}.W_C+b}]_j$.

In this model, matrix W_c and vector b are unknown parameters which should be learned in the training phase. Using error back propagation, these parameters are estimated. In training phase, the observed probability of occurrence of each single word is approximated for a given set of documents as follows:

$$P_{observed}(\boldsymbol{c}|w) = \frac{tf(\boldsymbol{c},w)}{tf(w)} \quad (6)$$

in which $tf(c,w)$ denotes the term frequency of w in documents dubbed as c, and $tf(w)$ represents the same in the entire collection of training set. Then, in the model construction step, we optimize the cross entropy of $H(p_{WE}, P_{observed})$ with *batch gradient decent* using Eq. (7).

$$L(W_C,b) = \frac{1}{m}\sum_{i=1}^{m} H(P_{WE},P_{observed}) + \frac{\lambda}{2m}\left(\sum_{i,j} W^2_{C_{i,j}}\right)$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{|c|} P_{observed}(c_j|w_i)\log P_{WE}(c_j|w_i)\right)$$
$$+\frac{\lambda}{2m}\left(\sum_{i,j} W^2_{C_{i,j}}\right) \quad (7)$$

in which $L(W_c,b)$ is the loss function, m denotes the size a training batch, and λ is the weight regularization parameter. The update rule for a particular parameter $\theta(W_C,b)$ given a single batch of size m is:

$$\theta^{(t+a)} = \theta^{(t)} - \alpha^{(t)} \odot \frac{\partial L(W_C^{(t)}, b^{(t)})}{\partial\theta} \quad (8)$$

$P(c|w)$ can be estimated for a pair of word w and class c using Eq. (5) after we train matrix W_c, and vector b. For the purpose of finding probability of each word for a given class c, Bayes' theorem can be used to estimate $P(w|c)$ in which it is approximated to be equal to $P(w|c) \approx P(w).P(c|w)$, where $P(w)$ denotes the prior probability of word w to be chosen as an important word. It has to be noted that we estimate $P(w)$ using TF-IDF calculated over the entire collection.

[7] Latent Dirichlet Allocation.

[8] In detection of viral messages problem, $x = 2$ (viral/non-viral classes) In category detection of viral messages, $x = 24$ (number of topics) In sentiment detection problem, $x = 3$ (number of sentiments including positive, negative, and neutral).

Table 5
Performance of baseline and proposed models.

Method	Viral Message Detection			Message Categorization			Sentiment Detection		
	P	R	F	P	R	F	P	R	F
NB	0.747	0.745	0.746	0.354	0.282	0.314	0.608	0.612	0.610
MI/LR	0.816	0.816	0.816	0.513[a]	0.281	0.363[a]	0.756[a]	0.622	0.682[a]
MI/SVM	0.805	0.778	0.791	0.469[a]	0.233	0.311	0.788[a]	0.584	0.671
EM/LR	0.839[a]	0.835[a]	0.837[a]	**0.527**[a]	0.343[a]	0.416[a]	0.758[a]	0.677[a]	0.715[a]
EM/SVM	0.762	0.753	0.757	0.448[a]	0.284	0.348[a]	0.730[a]	0.718[a]	0.724[a]
WE	**0.886**[a]	**0.885**[a]	**0.885**[a]	0.508[a]	**0.359**[a]	**0.421**[a]	**0.823**[a]	**0.827**[a]	**0.825**[a]

[a]Means statistically significant improvement over NB baseline.

Table 6
Top discriminative words in several categories.

Topic	Sport Events	Advertisements	Celebrity Artists
Word 1	International	Last	Said
Word 2	Team	Link	Music
Word 3	Goal	Channel	Actor/Actress
Word 4	Match	News	Reaction
Word 5	Cup	More	Singer

After computing the relevance probability of a word to a given class using either of methods above, we can estimate the probability of relevancy of message msg to a given class c as follows:

$$P(c|msg) = \frac{P(c) \times P(msg|c)}{P(msg)} \tag{9}$$

where $P(c)$ and $P(msg)$ are the prior probabilities of class c and message msg respectively and are assumed uniform in this paper. Thus, Eq. (9) is simplified to: $P(c|msg) \approx P(msg|c)$. Furthermore, assuming that message msg is composed of words $w_1, w_2, w_3, \ldots, w_n$ and each word w is generated independently, the probability of a message given the class is calculated as follows:

$$P(msg|c) = \prod_i P(w_i|c). \tag{10}$$

where $P(w_i|c)$ is estimated using either of the proposed methods.

7. Experiments

In this section, we first introduce the parameters of models and settings for the experiments. Then, we analyze the results.

7.1. Experimental setup and parameter setting

First of all, the parameters of the proposed models, and learning and testing process of classifiers are discussed.

For training and testing mentioned classifiers in Section 6.1, five fold cross validation has been employed to evaluate the results. In WE method, we have used adadelta ($\rho = 0.95$, $\epsilon = 10^{-6}$) with batch gradient descent and weight decay $\lambda = 0.01$ to optimize the loss function. The number of topics in LDA is set to 100. Additionally, we utilized Tensorflow to calculate matrix operations on a Nvidia Titan X GPU. When running SVM classifier, Radial basis function kernel with $\gamma = 0.1$ has been used. Ultimately, we ran EM algorithm with $\lambda = 0.1$ and 10 iterations.

7.2. Experimental results and analysis

Table 5 indicates the comparison of Naïve Bayes (NB) which is selected as baseline and the proposed classifiers and Word Embedding (WE) approach. According to this table, Precision (P), Recall (R) and F-Measure (F) of viral message detection problem for all approaches are high. This means that there exists some specific words which can distinguish viral from non-viral words which we explain them in

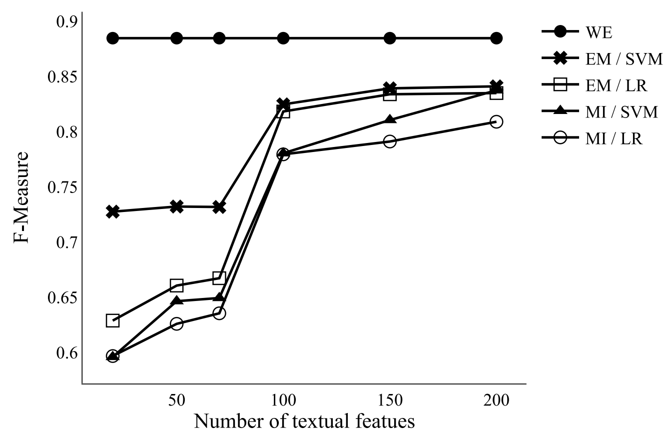

Fig. 7. Number of features impact on models performance.

Section 7.2.1 and also in Table 6. In this problem, WE method performs better than statistical classifiers. In comparison of EM and MI feature selection methods, we found that almost in all cases, EM approach outperforms owing to the fact that EM is a word clustering approach that promotes words which are more relevant to the target category and meanwhile are far from other categories. It is worth mentioning a two-tailed paired t-test ($\alpha = 0.05$) has been employed to measure statistical significance that is illustrated by star mark on the table.

7.2.1. Textual evidence - RQ2

In this section, we first study the effect that the number of words have on the classification problem and then, we will present top words for several sample categories of messages. Fig. 7 Indicates the F-Measure of proposed methods for different numbers of textual features for viral message detection problem. As illustrated in this figure, while increasing the number of textual features can considerably boost the performance of the model when less than 100 features are employed, after 100 features the performance will not change significantly. Furthermore, for all features, WE method outperforms statistical methods.

Table 6 demonstrates top discriminative words extracted by EM approach for three sample categories of messages. Since these words are not in English in the original dataset, we have carefully translated them to their equivalents in English as well as omitting stop words and proper nouns. As presented in this table, the words extracted for each category are ones that are relevant to that category and commonly used while writing about that topic. For instance, many of the advertisement messages sent in channels contain a "link" to get "more" information regarding that subject. Accordingly, highlighted words (i.e. *link* and *more*) are among the most important words for that category, extracted by EM approach.

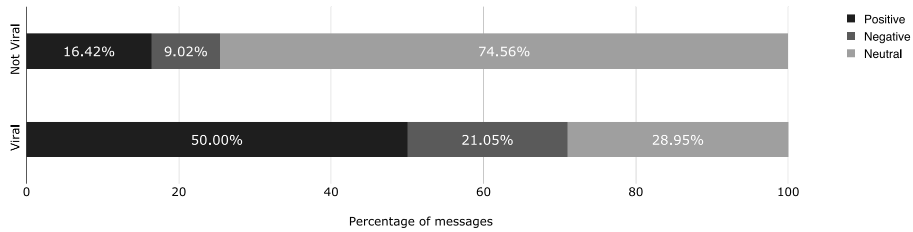

Fig. 8. Sentiment of the viral and non-viral messages in a sample channel.

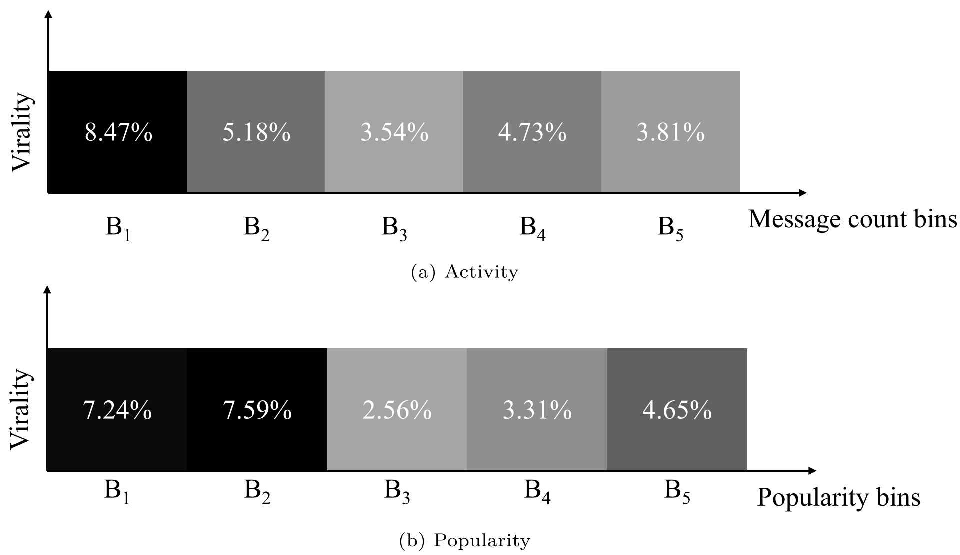

Fig. 9. The effect of channels popularity and activity on virality ratio.

7.2.2. The effect of sentiment on message virality - RQ3

In order to study the effect of messages sentiment and their virality, we have selected a general news channel and annotated its messages. The channel has published 752 messages during the time period covered in our dataset and among them, 76 were viral ones. The results are summarized in Fig. 8. As shown, near 75% of non-viral messages do not imply any sentiment, whereas, only 29% of viral messages do not carry a sentiment. This observation, in addition to other experiments we have applied on the dataset, suggests that messages with either a positive or negative sentiment are more probable to become viral. To analyze this observation from a statistical point of view, we have utilized Pearson's χ^2 test (Han et al., 2011) with one degree of freedom. Accordingly, considering that we have 76 viral messages among which 22 are neutral (i.e. carry no sentiment) and 676 non-viral ones containing 504 neutral messages, χ^2 value would be 67.61 which indicates a strong correlation, even with $\alpha = 0.001$ significance level, between sentiment and virality of messages.

7.2.3. The effect of channels' activity and popularity on message virality - RQ4

In this section, we aim to study how the number of messages published by the channels (i.e. activity) and their subscribers' count (i.e. popularity) may affect virality of their messages. Accordingly, we first define *virality ratio* $= \frac{\text{Number of viral messages sent by the channel}}{\text{Number of all messages published in the channel}}$ as a comparable metric that measures to what extent messages of a channel have become viral. Fig. 9 illustrates virality ratio for the channels in our dataset. In this figure, the channels are clustered into 5 bins with regard to their activity and popularity. To be more specific, B_1 includes the quintile of messages with the lowest activity/popularity and B_5 is the quintile of messages having the highest value of activity/popularity. Moreover, the darkness of the cells indicates virality ratio of the channels clustered in that cell.

As shown in Fig. 9(a), messages sent in the channels which publish fewer number of messages (i.e. bin B_1), are more probable to become viral. This can be explained by the behavior mentioned in Section 4.3. Accordingly, when a channel publishes a large number of messages every day, many subscribers would skip them and, as a result, messages in those channels have less opportunity to be forwarded by many users and become viral regardless of their content.

Since the exact number of channels' subscribers is not available in our dataset, we have exploited the average view count of each channel as a heuristic to form up the bins in Fig. 9(b). Two observations can be made from this figure. First, channels with superior number of subscribers (i.e. bin B_5) tend to have more viral messages due to the larger number of users who are reading their messages and may forward them. These channels are mostly well-known news agencies and celebrities' channels. Interestingly, on the other hand, channels with fewer number of members (i.e. bins B_1 and B_2) have the largest virality ratio. Generally, these channels do not belong to agencies and celebrities that are known by people. As a result, users would subscribe to such channels when they trust the content provided by that channels and accordingly, they may forward more messages from these channels that increases the virality ratio for less-popular channels.

8. Conclusion and future work

In this paper, we introduced Telegram IM Application and then compared it with Twitter and WhatsApp networks. Our main finding here is that Telegram network is a unique social network which brings both public and private messaging into a single environment. We investigated the information flow of public messages into groups and private chats and accordingly, defined the notion of viral messages. In fact, our main research question is to find potential viral messages. By identifying the characteristics of these messages, we will be able to predict them. Statistical classification and word embedding approaches are proposed to detect this kind of messages, their category, and sentiment. Our approaches extract the most discriminative words for each class of messages and use them as an important feature for the classification task. Our observations show that the messages with either a positive or negative sentiment are more probable to become viral. On the other hand, regarding the relationship between the channels' activity and the message virality, we found that when a channel publishes a large number of messages every day, many subscribers would skip them. Thus, messages in those channels have less opportunity to become viral regardless of their content.

Future work should target the privacy aspect of information flow in Telegram, as well as channel recommendation based on each user's preferences. Furthermore, using NLP-based methods on sentiment analysis may help to improve the results. On the other hand, the annotations of each message (i.e. sentiment and category) can be used as classification features for virality detection. Given that this paper shows that Telegram has special properties compared to Twitter and WhatsApp, it is necessary to study its different aspects in terms of text and network (e.g. the graph of original messages and messages forwarded to other channels). Although our definition of viral content is fine-tuned for the Telegram network, it can be simply generalized by slight changes for other social networks. For example, the number of re-tweets for a tweet sent by a specific user can be compared with other neighboring tweets by the same user to detect viral messages in that platform. Therefore, investigating viral contents on Twitter using the same method described in this work can be an interesting future work.

CRediT authorship contribution statement

Arash Dargahi Nobari: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Malikeh Haj Khan Mirzaye Sarraf:** Data curation, Writing - original draft, Methodology, Software. **Mahmood Neshati:** Supervision, Writing - review & editing, Investigation, Formal analysis, Conceptualization. **Farnaz Erfanian Daneshvar:** Visualization, Software, Writing - original draft, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

No funding was received for this work.

Intellectual property

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

References

Almerekhi, H., Hasanain, M., & Elsayed, T. (2016). EveTAR: A new test collection for event detection in Arabic tweets. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 689–692).

Alshammari, A., Kapetanakis, S., Evans, R., Polatidis, N., & Alshammari, G. (2020). User modeling on Twitter with exploiting explicit relationships for personalized recommendations. In *Hybrid intelligent systems* (pp. 135–145).

Anagnostopoulos, A., Petroni, F., & Sorella, M. (2018). Targeted interest-driven advertising in cities using Twitter. *Data Mining and Knowledge Discovery*, 737–763.

Anglano, C., Canonico, M., & Guazzone, M. (2017). Forensic analysis of Telegram Messenger on Android smartphones. *Digital Investigation*, 31–49.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 1130–1132.

Camarero, C., & José, R. S. (2011). Social and attitudinal determinants of viral marketing dynamics. *Computers in Human Behavior*, 2292–2300.

Cha, M., Benevenuto, F., Haddadi, H., & Gummadi, K. (2012). The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 991–998.

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. D., Stringhini, G., & Vakali, A. (2017). Detecting aggressors and bullies on Twitter. In *Proceedings of the 26th international conference on world wide web* (pp. 767–768).

Church, K., & de Oliveira, R. (2013). What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the 15th international conference on human-computer interaction with mobile devices and services* (pp. 352–361).

Dargahi Nobari, A., Reshadatmand, N., & Neshati, M. (2017). Analysis of telegram, an instant messaging service. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 2035–2038).

Dargahi Nobari, A., Sotudeh Gharebagh, S., & Neshati, M. (2017). Skill translation models in expert finding. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 1057–1060).

Doulamis, N. D., Doulamis, A. D., Kokkinos, P., & Varvarigos, E. M. (2016). Event detection in Twitter microblogging. *IEEE Transactions on Cybernetics*, 2810–2824.

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: Almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, 127–137.

Ghaffari, M., Rakhshanderou, S., Mehrabi, Y., & Tizvir, A. (2017). Using social network of TELEGRAM for education on continued breastfeeding and complementary feeding of children among mothers. *International Journal of Pediatrics*, 5275–5286.

Giachanou, A., Mele, I., & Crestani, F. (2016). Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 2263–2268).

Gregorio, J., Gardel, A., & Alarcos, B. (2017). Forensic analysis of telegram messenger for windows phone. *Digital Investigation*, 88–106.

Gupta, S., Khattar, A., Gogia, A., Kumaraguru, P., & Chakraborty, T. (2018). Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach. In *Proceedings of the 2018 world wide web conference* (pp. 529–538).

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hasan, M., Orgun, M. A., & Schwitter, R. (2018). A survey on real-time event detection from the Twitter data stream. *Journal of Information Science*, 443–463.

Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, 1146–1165.

Hua, T., Chen, F., Zhao, L., Lu, C.-T., & Ramakrishnan, N. (2013). STED: semi-supervised targeted-interest event detection in twitter. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1466–1469).

Jalilvand, A., & Neshati, M. (2020). Channel retrieval: finding relevant broadcasters on Telegram. *Social Network Analysis and Mining, 10*, 1–16.

Jiménez-Bravo, D. M., De Paz, J. F., & Villarrubia, G. (2019). Twitter's experts recommendation system based on user content. In *Distributed computing and artificial intelligence, special sessions, 15th international conference* (pp. 251–258).

Keib, K., Himelboim, I., & Han, J. (2018). Important tweets matter: Predicting retweets in the #BlackLivesMatter talk on twitter. *Computers in Human Behavior*, 106–115.

Kim, Y., & Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, 59–77.

Li, Q., Nourbakhsh, A., Shah, S., & Liu, X. (2017). Real-time novel event detection from social media. In *2017 IEEE 33rd international conference on data engineering* (pp. 1129–1139).

de Lira, V. M., Macdonald, C., Ounis, I., Perego, R., Renso, C., & Times, V. C. (2019). Event attendance classification in social media. *Information Processing & Management*, 687–703.

Mele, I., Bahrainian, S. A., & Crestani, F. (2019). Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management*, 969–993.

Negreira-Rey, M.-C., López-García, X., & Lozano-Aguiar, L. (2017). Instant messaging networks as a new channel to spread the news: Use of whatsapp and telegram in the spanish online media of proximity. In *WorldCIST, Advances in intelligent systems and computing* (pp. 64–72).

de Oliveira, J., Santos, D. H., & Neto, M. P. (2016). Chatting with arduino platform through telegram bot. In *2016 IEEE international symposium on consumer electronics* (pp. 131–132).

Paltoglou, G. (2016). Sentiment-based event detection in Twitter. *Journal of the Association for Information Science and Technology*, 1576–1587.

Piwek, L., & Joinson, A. (2016). What do they snapchat about? Patterns of use in time-limited instant messaging service. *Computers in Human Behavior*, 358–367.

Qiu, J., Li, Y., Tang, J., Lu, Z., Ye, H., Chen, B., Yang, Q., & Hopcroft, J. E. (2016). The lifecycle and cascade of wechat social messaging groups. In *Proceedings of the 25th international conference on world wide web* (pp. 311–320).

Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., & Schmidt, A. (2014). Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3055–3064).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860).

Schouten, K., van der Weijde, O., Frasincar, F., & Dekker, R. (2018). Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE Transactions on Cybernetics*, 1263–1275.

Seufert, M., Schwind, A., Hoßfeld, T., & Tran-Gia, P. (2015). Analysis of group-based communication in whatsapp. In *Mobile networks and management* (pp. 225–238).

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 959–962).

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268.

Sušánka, T., & Kokeš, J. (2017). Security analysis of the telegram IM. In *Proceedings of the 1st reversing and offensive-oriented trends symposium* (pp. 1–8).

Telegram (2020). Telegram official blog. https://telegram.org/blog, (Accessed 24 Apr 2020).

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 2544–2558.

Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y., & Li, T. (2015). Dual sentiment analysis: Considering two sides of one review. *IEEE Transactions on Knowledge and Data Engineering*, 2120–2133.

"Xiaomei, Z., Jing, Y., & Jianpei, Z. (2018). Sentiment-based and hashtag-based Chinese online bursty event detection. *Multimedia Tools and Applications*, 21725–21750.

Yang, J., Qiao, Y., Zhang, X., He, H., Liu, F., & Cheng, G. (2015). Characterizing user behavior in mobile internet. *IEEE Transactions on Emerging Topics in Computing*, 95–106.

Yuan, K., Liu, G., & Wu, J. (2019). Whose posts to read: Finding social sensors for effective information acquisition. *Information Processing & Management*, 1204–1219.

Yusoff, M. N., Dehghantanha, A., & Mahmod, R. (2017). Network traffic forensics on firefox mobile OS: Facebook, Twitter, and telegram as case studies. *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, 63–78.

Zainuddin, N., Selamat, A., & Ibrahim, R. (2018). Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1218–1232.