

Sep_TD_Tel01

Mehrdad Ranjbar-Khadivi^{1,2}, Mohammad-Reza Feizi-Derakhshi¹, Aynaz Forouzandeh¹, Pejman Gholami¹, Ali-Reza Feizi-Derakhshi¹, and Elnaz Zafarrani-Moattar^{1,3}

¹Computerized Intelligence Systems Laboratory, Department of Computer Engineering, University of Tabriz, Tabriz, Iran

²Department of Computer Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran

³Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

January 6, 2022

1 Introduction

The Sep_TD_Tel01 dataset was compiled by ComInSyS¹. This database has been compiled due to the low resource of the Persian language and the high popularity of the Telegram social network in Iran. In this work, an official API published by Telegram is used. To respect the principle of privacy, only data related to public channels and public groups have been collected.

This dataset contains 10,209 records of messages sent to public channels and groups in the one month between 1 January 2017 (12 Day 1395) and 31 January 2017 (12 Bahman 95). This database is divided into sixty 12-hour windows, which include two super-hot topics: "The death of Ayatollah Hashemi Rafsanjani" and the second topic, "Plasco building fire". To be able to review and control the performance, nine of these windows have been selected as GT and labeled. These windows are [14, 15, 16, 17, 18, 37, 38, 39, 40]. A summary of information about this database is given in the table below.

Number of posts	10209
Number of super-Topics	2
Number of sub-Topics	81
Number of windows	60
Number of GT windows	9
The average number of each window post	170
Size of each window (hours)	12
Number of labeled posts	2365

Table 1: summary of Dataset.

2 Data structure

The general structure of the database is given below.

¹cominsys.ir

2.1 posts

Each message on the Telegram social network is recorded in this table. It should be noted that the topic ID starts from one and the "0" Id in this table indicates the non-topic title and each message with a "0" topic shows that the message does not contain an important topic that has newsworthy.

Name	Type	Null/Ref	Description
Seq	Long int		Sequence number
Chat Name	String		Channel/Group name
DateSent	Date/Time		The message's sent time
Text	string		Post's text
WindowNum	int		Windows number of Post
Topics	list	Null	Topics' Id

Table 2: Posts

2.2 Topics

Each of the extracted topics (labels) is stored in this table

Name	Type	Null/Ref	Description
Id	Long int		identity
Topic	string		Topic title

Table 3: Topics

3 Data storage

This database is presented in two file types, Excel and JSON, both of which have the same content.