TempTracker: A Service Oriented Temporal Natural Language Processing Based Tool for Document Data Characterization and Social Network Analysis

Onur Can Sert
Information Technologies, Zalando, Ireland Ltd
onur.can.sert@zalando.ie

Sibel Tariyan Özyer
Department of Computer Engineering, Ankara Medipol University, Turkey
sibel.tariyan@ankaramedipol.edu.tr

Deniz Bestepe
Department of Computer Engineering, Istanbul Medipol University, Turkey
deniz.bestepe@medipol.com.tr

Tansel Özyer
Department of Computer Engineering, Ankara Medipol University, Turkey
tansel.ozyer@ankaramedipol.edu.tr

Abstract: *With the advent of Web 2.0 based technology, news sites and micro-blog sites have become popular and have attracted the attention of people around the world. Existing textual data captured by these sites is highly beneficial for extracting (a) new information to analyze, and (b) temporal course of change in entities, topics and sentiment for differing granularities. This has been demonstrated by the study described in this paper. After collecting the data, several directions have been investigated in order to demonstrate its effectiveness under the umbrella of entity extraction, topic and sentiment analysis using Natural Language Processing (NLP) tools, temporal social media analysis, and time varying trend results of entity and sentiment aspect of entities. A service-based architecture has been proposed to process text data with NLP tools and to enrich the data. Text data is collected and processed via NLP tools. It is retrieved upon request for data analysis. The reported results illustrate the applicability and effectiveness of the conducted study.*

Keywords: *Natural language processing, entity recognition, sentiment analysis, social network analysis.*

Received August 7, 2021; accepted October 7, 2021
https://doi.org/10.34028/iajit/19/3/8

1. Introduction

In the digital age, high proportion of people foster information sources such as online news, social networking and micro-blogging sites. A variety of available interpretation techniques may be utilized to better benefit from the captured data. This forms the basis for enrichment leading to effective knowledge discovery. Computational linguistics, also known as natural language processing techniques, are utilized for learning, understanding and producing human language content. Some classical examples of Natural Language Processing (NLP) are speech recognition, machine translation, document summarizing and spoken dialogue systems, sentiment analysis, named entity recognition, Part-Of-Speech (POS) tagging, morphological analysis organization, among others [10].

Several toolkits have been proposed to accomplish NLP tasks, including Apache OpenNLP [2], Stanford CoreNLP [13] and Opener [1]. These provide operations such as tokenizer, sentence splitter, pos tagger, named entity recognition, sentiment analysis, Co-reference resolution, etc., Studies suggest OpenNLP as the best choice for news [20]; and voting based ensemble approach outperforms OpenNLP toolkit [3]. The study described in this paper builds on some similar studies described in the literature. For instance, the study described in [11] performs clustering of individual publications and aggregated levels. Apache OpenNLP is used for post-tagging and irrelevant words are filtered out. Relevant phrases are identified, mapped and visualized. Finally, a clustering method maximizes a quality function based on relatedness [11].

A RESTful service-based text annotation tool was implemented in 2011 [22]. In started with the annotation of six different corpora by four research groups. It ended up with the creation of over 50,000 annotations in thousands of lengthy documents. NLP functions such as POS-tagged tokens, chunking, entity and dependency annotation were utilized [22]. Net-Builder [6] is a tool to investigate and reveal hard to discover characteristics of terrorist networks. It constructs a social network of actors and predicts the

links between terrorists. Our study adopts a similar way of network construction and visualization by using social network analysis in the same manner [6].

Another study takes stack overflow website, extracts Latent Drichlet Allocation (LDA) based topics, and uses the extracted topics to identify overlapping communities of users for the question and answering system. One limitation of the system is the need to give tag sets with the question [14, 15]. NLP Toolkits are used to handle named entity recognition [19], sentiment analysis [18] and topic extraction [4] utilities. They have become the basic elements for analyzing the change within a group over time. They may also be extended longitudinally from a social network analysis perspective [24]. Another direction is to find regularities in data over time; this has been left as future work where we plan to use association rule mining which is an active area in data mining to find non-trivial, hidden, sophisticated information. Although the above-mentioned techniques have been applied to online news data, they can be integrated differently to produce more effective outcome.

The study described in this paper proposes a service-based framework which contains different natural language processing toolkits and algorithms, gives opportunity to create custom pipelines with using selected items and configurations to make different types of analysis. In addition to running different natural language processing algorithms, the proposed framework also introduces different functionalities to build social networks and make analysis on the created networks with using the results that are created by the natural language processing pipeline. The benefit of using a service-based approach is to overcome cross-platform integration difficulty and make it augmentable with different NLP toolkits. The contributions of the study described in this paper may be listed as follows:

- A service based extensible framework for data analysis.
- Extraction of entities, and topics with or without sentiment orientation using NLP toolkits.
- Retrieval of statistical summary of information at different time intervals (year, month, day).
- Network construction from a summary result and a social network analysis for subsequent time intervals.
- Comparative trend analysis of different elements for monitoring.

The outline of the paper is as follows. The necessary background is covered in section 2. The service architecture utilized in this study is described in section 3. The analysis and the results are reported in section 4. Section 5 is conclusions.

2. The Necessary Background

2.1. NLP Toolkits

In this study, we have incorporated three different NLP libraries, namely Apache OpenNLP, Stanford CoreNLP and Opener. Apache OpenNLP is a Java based library. It is a machine learning based toolkit for processing text. It can be used for both academic and industrial purposes [2]. It primarily contains fundamental modules such As Part Of Speech (POS) Tagging, Named Entity Recognition (NER), tokenization, language detector, sentence detector, document categorizer, and parser. Furthermore, the toolkit is also used for machine learning based applications. Stanford CoreNLP [13] has been developed by Stanford University for linguistic analysis in textual structures. It has been implemented in Java. It is open source. Furthermore, Stanford CoreNLP can be easily integrated with different programming languages, including Python, Ruby, Perl, Scala, JavaScript,. NET and C#. It supports a variety of languages, including Arabic, Chinese, English, French, German and Spanish. Open Polarity Enhanced Named Entity Recognition (OPENER) [1] is European Commission funded project under the 7th Framework Program. It is based on Cloud service architecture. It has various NLP modules, such as language identifier tokenizer, tree tagger, polarity tagger, property tagger, constituent parser, kaf-naf parser, tokenizer, tree tagger, POS tagger, polarity tagger, property tagger, constituent parser, kaf-naf parser, named entity recognition, scorer, and opinion detector. Results are generated in Knowledge Annotation Framework (KAF) format. Later, it can be transformed into Extensible Markup Language (XML). It supports languages like English, Spanish, German, French, Italian and Dutch. A comparison of these toolkits is summarized in Table 1.

Table 1. Tools and programs (1) OpeNER (2) Apache OpenNLP, (3) Stanford CoreNLP.

Features and Components	1	2	3
Language Identifier	X	X	X
Tokenizer	X	X	X
POS Tagger	X	X	X
NER	X	X	X
Named Entity Linking	X		
Sentiment Analysis	X	X	X
Co-reference Resolution	X	X	X
Bootstrapped Pattern Learning			X
Open Information Extraction			X
The Parser	X	X	X

2.2. Basic Functions of NLP

2.2.1. POS (Part-Of-Speech) Tagger

POS Tagger is a module which reads some text and matches the words with special morphological tags, such as noun, verb, adjective, etc., the text is tokenized into words and tags are assigned to words. The POS

Tagger module contains for each language different libraries which should be added to corresponding applications. Here, it is worth mentioning that the tag dictionary will be different in our NLP tools.

OpeNER POS Tagger is also based on Apache OpenNLP library. The component supports a variety of human languages. When using POS Tagger, the component uses 3 different modules for English and 2 for other supported languages. These modules are: Plain text dictionary, Using Morphologic-stemming and WordNet. The WordNet module is only used for English. The OpeNER's data input and output are different from those available in Apache OpenNLP and Stanford CoreNLP. Input and output types are KAF [1]. Pos Tagger tag types are listed in Table 2 [23].

Table 2. POS tagger tags [23].

CC Coordinating conjunction	TO Infinitivial to
CD Cardinal number	UH Interjection
DT Determiner	VB Verb, base form
EX Existential there	VBD Verb, past tense
FW Foreign word	VBG Verb, gerund or present participle
IN Preposition or subordinating conjunction	VBN Verb, past participle
JJ Adjective	VBP Verb, non-3rd person singular present
JJR Adjective, comparative	VBZ Verb, 3rd person singular present
JJS Adjective, superlative	WDT Wh-determiner
LS List item marker	WP Wh-pronoun
MD Modal	WP$ Possessive wh-pronoun
NN Noun, singular or mass	WRB Wh-adverb
NNS Noun, plural	# Pound sign
NNP Proper noun, singular	$ Dollar sign
NNPS Proper noun, plural	. Sentence-final punctuation
PDT Predeterminer	, Comma
POS Possessive ending	: Colon, semi-colon
PRP Personal pronoun	(Left bracket character
PRP$ Possessive pronoun	) Right bracket character
RB Adverb	"Straight double quote
RBR Adverb, comparative	' Left open single quote
RBS Adverb, superlative	"Left open double quote
RP Particle	' Right close single quote
SYM Symbol	"Right close double quote

2.2.2. Named Entity Recognition (NER)

Named Entity Recognition is a task that aims to find, classify and label named entities in the given input. Labels that can be identified with named entity recognition algorithms can be location, person, organization, association, country, date and others [19]. Stanford CoreNLP Named Entity Recognition has three different tag class models (person, location, organization). NER types are determined by Conditional Random Fields (CRF) sequence taggers and predefined rules for some cases.

Apache OpenNLP uses seven different NER types. They are Person, Location, Date, Money, Organization, Percentage and Time. Tags are customized and optionally used. There exist predefined models, for example, en-nerdate.bn, en-ner-location.bin, en-ner-organization.bin, en-ner-person.bin, anden-ner-time.bin to be able to perform NER type identification in Java applications. In OpeNER, NER types are Location, Person, Organization, Money, Percent, Date, Time.

2.2.3. Sentiment Analysis

Sentiment Analysis is a process which determines the category of a review or text (e.g., positive or negative) [18]. Document categorizer in apache OpenNLP carries out classification according to the available categories. In order to use Document Categorizer, en-doccat. Bin file is needed for English language. Sentiment Analysis with Stanford CoreNLP implements Socher's sentiment model. Stanford CoreNLP Sentiment Analysis determines classes by using deep learning techniques. It has five different classes such as very positive, positive, neutral, negative and very negative. OpeNER has positive and negative sentiment results which can be associated with named entities. Rather than acquiring the polarity of the entire content, it may be preferred to perform subtask as mentioned in [8] to decompose the entire message in order to understand the role of entities contextually as positive or negative. A previous study described in [7] has been adapted to twitter components to obtain the ensemble of different aspects such as (message, emoticons, word2vec, negations, etc.,) with Support Vector Machine (SVM) classification [16]. A different study also shows that named entity recognizers can be trained and used for different cases and domains such us software development environment [12].

2.2.4. Topic Modelling

Latent Dirichlet Allocation (LDA) algorithm is a very popular unsupervised learning approach to discover topics. There are different implementations of LDA topic modelling algorithm in different programming languages. LDA is a probabilistic model which works with text corpus.

Corpus is a collection of documents. Let's assume we have n different documents as $D=M1, M2,....., Mn$. In this case LDA processes each document in set D and produces an output for every document. Output refers to topics and LDA is generally used for text modeling. This model is based on three-levels Bayesian. LDA calculates corpus probability [4]. Further, topic coherence model [21] is used to determine the number of topics.

2.3. Social Network Analysis

A social network bears a structure made up of nodes with various types and their connections may show different types of relationships such as interest, like, proximity, etc., Social network analysis adapts graph-based structures. There are also additional concepts such as dyad, triad, relational tie to represent connections within groups and subgroups. There are

different kinds of edge types (relations) [24]. Social network analysis also refers to the process of change within time intervals. For finding patterns within time intervals different strategies can be followed [9]. An overview of social network analysis tools is covered [5].

3. System Architecture

In our study, we designed a scalable and flexible system architecture. Modules located in the system are built for collecting data from different data sources, applying natural language processing techniques in different toolkits, and analysing the results. The overall system architecture can be seen in Figure 1. According to the architecture given in Figure 1, core of the system is designed as a web service. As a result of that design, this system can be easily enriched with new modules according to the requirements and can be used by different applications. The overall system split into 3 main parts. These parts are; Data Collection Services, NLP Services, and Data Analysis Services The overall system has been prepared by using different technologies both in the server side and client side. For building server side, we used mainly JavaScript, Java and Python based technologies. In database both relational database technologies and Not only SQL (NoSQL) technologies are used. According to the type and length of the data, it is stored in either MySQL database or Mongo database. For building the bridge between server side and client-side Personal Home Page (PHP) is used. In the client side, essential web technologies such as Hypertext Markup Language (HTML), JavaScript and Cascading Stylesheets (CSS) are used. Finally, Frequent Pattern (FP) Growth algorithm and Pajek tool [17] has been used for data analysis.

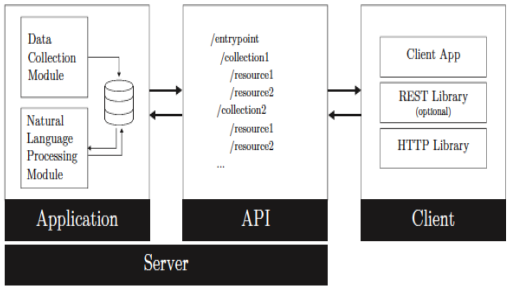

Figure 1. Overall system architecture.

3.1. Data Collection Services

To use NLP services, we need to have text data. We can use both online data and stored data. Huge amount of text data can be found in different sources in the internet. Such valuable data can be collected using a variety of techniques, e.g., web crawlers and Application Programming Interface Application Programming Interface (APIs). Web crawlers, also known as spiders, automatically navigate web pages and collect specified data from them. APIs are offcial web services released by companies or websites. They are more reliable structures compared to web crawlers. However, APIs are source specific structures developed by the owner of the data source.

In our study, we developed two different data collection services using these techniques for collecting different types of text data from various sources. These services are developed using JavaScript technologies. The structure of the data collection service can be seen in Figure 2.

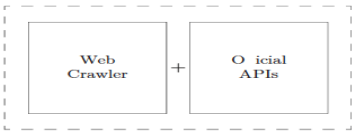

Figure 2. Overview of the data collection services.

3.2. NLP Services

For mining the text data collected from the services or stored in our database, we have built a natural language processing service. This service obtains text data and processes them with NLP toolkits for NER, sentiment and LDA output. NLP toolkits output results based on majority voting [3]. The structure of the NLP Service is shown in Figure 3.

Apache OpenNLP, Stanford CoreNLP, OpeNER toolkits have been used for building the core of the entity recognition and sentiment analysis services. Java releases of the toolkits are used for sustainability and easy integration. In addition, any other NLP toolkit can be added to the architecture. Further, Python release of the gensim library is used for the topic modelling service.

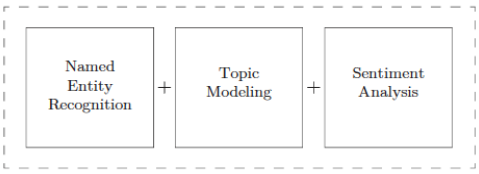

Figure 3. Overview of the NLP services.

3.2.1. Named Entity Recognition Service

In our study, to increase the success rate, we built a scalable service that includes different tools which employ a variety of methods to find named entities. This service gets text data as input, runs different tools in parallel, generates results and combines them. The structure of this service is shown in Figure 4.

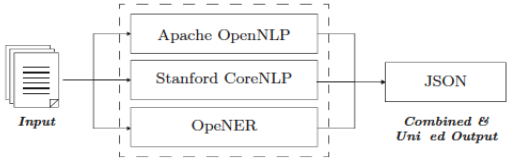

Figure 4. Structure of named entity recognition service.

According to our architecture, NLP services take text as input. However, each natural language tool gives output in a different format. To create a robust and user-friendly service, our system requires output in one format. We handled this challenge by developing a result unification module which takes results from different natural language processing tools and produces a unified result in JavaScript Object Notation (JSON) format. JSON is a frequently used data format, especially for web services. Because it is compatible with many different technologies, it is very flexible and easy to parse and convert to other data formats such as xml, csv, string, etc., It can also be easily stored in NoSQL databases such as MongoDB and CouchDB, as well as in Relational Database Systems like MySQL, PostgreSQL etc., Assume that the system accepts the following sentence as input to the named entity recognition service:

"... a critical crossing in the suburbs of New York City. Jamey Barbas, the engineer orchestrating the project for the Thruway Authority ..."

Apache OpenNLP lists all named entities, their types and indexes in a text format. Apache OpenNLP's example output format is shown in Figure 5.

```
[463..466) location New York City
[467..469) person Jamey Barbas
[477..479) organization Thruway Authority
```

Figure 5. An example of apache OpenNER output format.

```
a
critical
crossing
in
the
suburbs
of
New LOCATION
York LOCATION
City LOCATION
.
Jamey PERSON
Barbas PERSON
,
the
engineer
orchestrating
the
project
for
the
Thruway ORGANIZATION
Authority ORGANIZATION
```

Figure 6. An example of stanford CoreNLP output format.

Stanford CoreNLP gives all the words and punctuation characters placed in the input text in labelled format. Final entities should be mined from the text result. Stanford CoreNLP's example output format is shown in Figure 6. OpeNER has a more structured and unique output format. It gives KAF document as an output. KAF documents are specialized XML documents formatted by the OpeNER community. OpeNER's example output format is shown in Figure 7.

```
<entity eid="e1" type="location">
  <references>
    <!--New York City-->
    <span>
      <target id="t463" />
      <target id="t464" />
      <target id="t465" />
    </span>
  </references>
</entity>
<entity eid="e2" type="person">
  <references>
    <!--Jamey Barbas-->
    <span>
      <target id="t467" />
      <target id="t468" />
    </span>
  </references>
</entity>
<entity eid="e3" type="organizatio
  <references>
    <!--Thruway Authority-->
    <span>
      <target id="t477" />
      <target id="t478" />
    </span>
  </references>
</entity>
```

Figure 7. An example of OpeNER output format.

Our named entity recognition service finds named entities using the abovementioned tools, parses their results and combines the outcome in JSON format using the unification module. An example of this unified result structure is shown in Figure 8.

```
'namedEntities': [
 ...
 {
  'sentence': '... a critical crossing in the suburbs of New York City',
  'namedEntity': 'New York City',
  'category': 'LOCATION',
  'startIndex': 463,
  'endIndex': 465
 },
 {
  'sentence': 'Jamey Barbas, the engineer orchestrating the project for the
Thruway Authority ...',
  'namedEntity': 'Jamey Barbas',
  'category': 'PERSON',
  'startIndex': 467,
  'endIndex': 468
 },
 {
  'sentence': 'Jamey Barbas, the engineer orchestrating the project for the
Thruway Authority ...',
  'namedEntity': 'Thruway Authority',
  'category': 'ORGANIZATION',
  'startIndex': 477,
  'endIndex': 478
 },
 ...
]
```

Figure 8. An example of unified result format.

3.2.2. Sentiment Analysis Services

We also built a service for making sentiment analysis similar to the named entity recognition service. Once again, after running the algorithms for the text input separately, we combined their results into a unified item. We categorized the entire content or each sentence in the content as positive or negative. In addition to using this service, we can generate sentiment analysis results for named entities which are located in the input. Named entity sentiment analysis can be done in two ways; differentiating entities with positive and negative polarity (e.g., trump p and trump n for positive and negative, respectively) as well as signed networks. For the sake of brevity, we have given the first one. The structure of the module is shown in Figure 9.

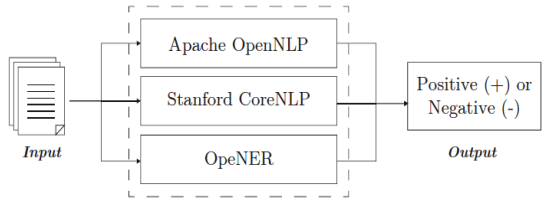

Figure 9. Structure of sentiment analysis service.

3.2.3. Topic Modeling Services

Unlike the named entity recognition service and the sentiment analysis service, we built a simpler service for topic modelling. We just used the latent dirichlet allocation algorithm in our module. The topic detection and coherence service has been implemented in Python using gensim library. It accepts a bulk of text content as input and generates two results; different topics and their keyword groups are generated from the input data and a content-topic matrix which stores topic scores for each text content. The structure of the module is shown in Figure 10.

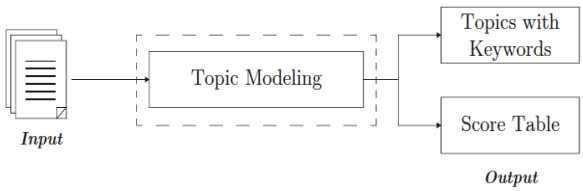

Figure 10. Structure of topic modeling service.

3.3. Data Analysis Services

Data analysis services are developed for analyzing the collected text data by applying natural language processing techniques. These services contain Pajek [17], which is a social network building and analysing tool and some algorithms such as association rules mining, hierarchical clustering, and graph layout algorithms. These services use named entity recognition, sentiment analysis and topic modelling outputs of NLP services and provide additional and more advanced analysis opportunities of the data.

4. Data Analysis

The client application carries out data analysis. News are filtered with date interval criteria. They are summarized at different scales temporally, including day, month and year. Named entities (Person, Organization, Location), topics and associated sentiment results are potential process results. After request, a client prepares the data sets for analysis. This section contains different analysis results for multiple perspectives. These analysis results show valuable connections and changes between different items.

Different values and metrics have been identified and compared during the analysis process. Two of those metrics are called emerging and disrupting. Emerging items represent entity pairs which started to appear together in the files and built a relationship in time. On the other hand, disrupting items represent entity pairs which lost their togetherness and destroyed their relationship in time according to the analysis. In addition to these metrics, we have also introduced metrics called increasing popularity and decreasing popularity. Increasing items represent the items which appear suddenly or become popular in a short period of time. In contrast, decreasing items represent the items which disappear from the results and lost their popularity in a short period of time.

4.1. NER Analysis

We have collected 12560 news articles of New York Times from January 2017 to December 2017 in four categories (headlines, subheads, international and national). We have analyzed named entities for the given time period and we have identified 41, 184 different entities in the data. When we analyzed them monthly, we have noticed that importance level of the entities is changing in time. To identify those entities and critical changes, we have listed the top 20 (the most critical) emerging and disrupting entities with their occurrence numbers in the news for the “united nations” entity and their statistical information.

We have selected “united nations” as a center point for the named entity recognition analysis because we saw that it has a quite broad coverage area for the named entities such as countries and political actors. In addition, we focused on political relations because of the data that we have and the entity values that we identified. We also analyzed the general and detailed forms of periodical changes in all emerging and disrupting entities.

4.1.1. Monthly Emerging and Disrupting Named Entities for the United Nations

Temporally stored network representations have been processed monthly for 2017 and only the ones linked to “united nations” entity have been considered further

to figure out the top 20 emerging and disrupting entities for the first 4 months of 2017 reported in Table 3. Besides, we have found transitions between months. While we were analyzing the transitions between months, we have focused on the most popular entities that we identified. To cover the most important entities, we have focused on the top 1% of the identified entities and collected the top 500 items. Then we have identified their statuses, such as emerging, disrupting, increasing popularity and decreasing popularity.

We have applied the hierarchical clustering method for the top 500 occurring entities. Also, with square root for the number of nodes threshold value, the corresponding hierarchical clustering results have been visualized in circular partitioned fashion. Each node may have different vertical border color to indicate emerge (orange), disrupt (black), increase (green) and decrease (blue).

Table 3. Top 20 emerged and disrupted named entities for "united nations" in the first 4 months of 2017.

Period 2017	Emerged	Disrupted
Jan to Feb	Mosul(193), Kushner(96), Qaddafi(84), Khalil(78), Rashidiya(72), Churkin(47), Sharif(42), Anastasiades(40), Quabba(36), Libyan(35), Malaysia(30), Tripoli(30),	Jammeh(227), Gambia(152), Barrow(91), Senegal(49), Juba(48), Central African Republic(36), Kiir(36), Rafsanjani(35), Machar(33), Paris(30), Puebla(30), Donald J. Trump(28), Wadi Barada(28), Hacking Team
Feb to Mar	Haiti(256), Khalaf(111), South Sudanese(70), Schumer(59), South Africa(58), Darfur(54), Kislyak(54), India(46), Yam-bio(42), Dubai(41), Liberia(38), Heritage Foundation(36), Mali(36), Groves(36), Moon(34), Levinson(34), Delattre(34), S.G.(32), Nepalese(30), Gayflor(28)	Kushner(96), Qaddafi(84), Khalil(78),Rashidiya(72), Myanmar(62), Churkin(47), Sharif(42), Anastasiades(40), Quabba(36), Libyan(35), Greek(33), Gaza(32), Tripoli(30), Baghdad(27), General McMaster(26), Al Bab(25), Unama(25), seven Muslim-majority(24), FARC(24), Nicosia(24), Mohamed Sharif(24),
Mar to Apr	Kosovo(157), Roma(98), Kony(56), Unmik(56), Aung San Suu Kyi(51), Resistance Army(49), Ugandan(35), Page(34), Zarrab(33), Khan Sheikhoun(32), Xi(29), Offce of Legal Affairs(28), Lombardi(27), Chechnya(23), Italy(22), Giuliani(20), Podobnyy(20),	Khalaf(111), apartheid(106), South Sudanese(70), Schumer(59), Netanyahu(57), Kislyak(54), Mosul(46), Flynn(44), Yambio(42), Dubai(41), Senate(39), Mali(36), Groves(36), Levinson(34), S.G.(32), Nigeria(31), Wang(31), Nepalese(30), Canada(29), Iraqi(29)

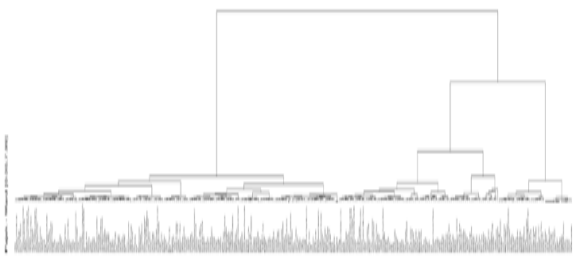

Figure 11. March to April entities with hierarchical clustering.

An example of hierarchical clustering of the named entities can be seen in Figure 11. This clustering result contains all the active entities during March 2017. In addition, the visualization of these clusters is shown in Figure 12.

Another example for a different month is shown in Figure 13. This clustering result contains all the active entities during April 2017. The visualization of these clusters can be seen in Figure 14. In the clustering result, it can be observed that some clusters continue to appear while some other clusters disappear over time.

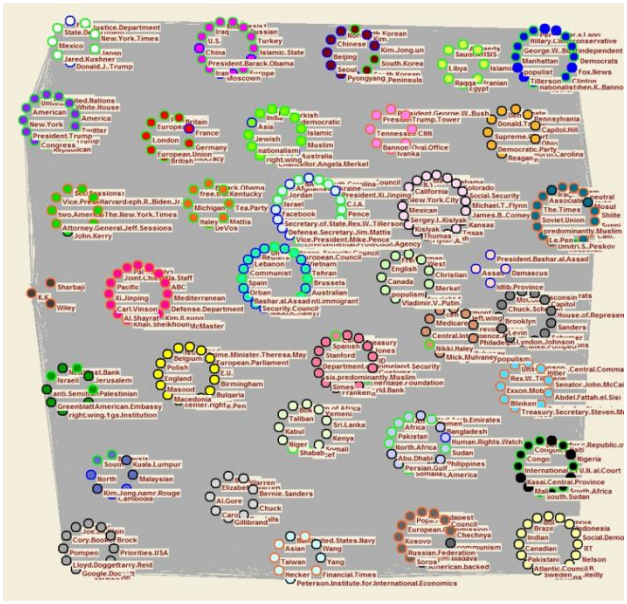

Figure 12. Named entity network created with most popular entities during March to April.

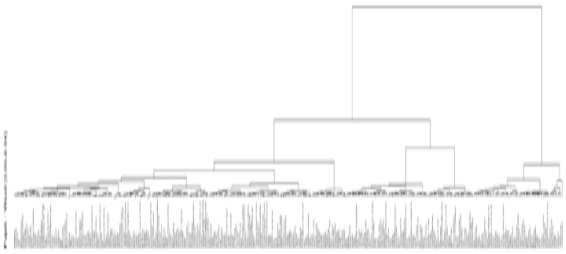

Figure 13. April to May entities with hierarchical clustering.

4.1.2. Counts of Monthly Emerging and Disrupting Named Entities for the

- United Nations Entity Case: the total numbers of emerged and disrupted entities are reported in Table 4. The lowest emerging number of entities is during March to June. It is April to July for the disrupted case. On the other hand, active periods are in August-September, January-February, June-July and November-December transitions for emerged, while active periods are September-October and February-March transitions for the disrupted case.

Table 4. Total number of named entities identified monthly around "united nations" item in 2017.

Period 2017	Emerged	Disrupted
Jan to Feb	793	580
Feb to Mar	704	758
Mar to Apr	522	721
Apr to May	525	559
May to Jun	548	532
Jun to Jul	742	534
Jul to Aug	643	718
Aug to Sep	849	594
Sep to Oct	613	859
Oct to Nov	604	640
Nov to Dec	712	619

4.1.3. The Most Significant Emerged and Disrupted Entities

The co-occurrences of entities which emerged and disrupted as the largest for each month are reported in Table 5. Also, Table 6 briefly discusses the importance of nodes by using eigenvector centrality in temporal networks. Table 7 details the monthly situation of the co-occurrences of the first 4 months of 2017 for those given in Table 5.

Table 5. The most significant emerged and disrupted named entities for the entire network.

Period 2017	Emerged	Disrupted
Jan to Feb	Trump-Flynn	Mexico-Nafta
Feb to March	Trump-Ryan	Trump Bannon
April to March	Trump-CNN	Trump-Ryan
May to April	Trump-Flynn	Fox News-O'reilly
May to June	Army- Manning	Aleppo-Assad
June to July	Trump-Putin	Army-Manning
July to August	Trump-Bannon	China-Liu
August to September	Trump-Democrats	Iran-Mullah Mansour
September to October	Spain-Catalonia	Trump-Bannon
October to November	Lebanon-Hariri	Spain-Catalonia
Nov to Dec	Trump-Jerusalem	Senate-House

Table 6. Association change for the entities of Table 5 for each month in 2017. January contains the initial value, other months have the change.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Trump-Flynn	0	3471	-2479	-992	2185	-947	-1238					
Mexico-Nafta	1262	-1262					68	-68				
Trump-Ryan	0		5375	-5375								
Trump-Bannon	0	1600	-1600	2791	-2791			6127	-4744	-1383		
Trump-CNN	0			10583	-10221	246	-608	337	572	-706	248	250
Fox News - O'Reilly	0			3792	-3792							
Army-Manning	0					2832	-2832					
Aleppo-Assad	33	55	-50	14	1772	-1824		14	-14		16	1
Trump-Putin	1256	-419	-284	479	198	-1230	4005	-3440	-451	207	1217	-1005
China-Liu	0					2390	-2390					
Trump-Democrats	4103	-339	3336	4022	918	-1664	-159	-2173	2803	-623	-208	576
Iran-Mullah Mansour	0							1368	-1368			
Spain-Catalonia	0									2783	-2783	620
Lebanon Hariri	0						13	-13			1755	-1040
Trump-Jerusalem	428	-243	-62	-123	730	-730	7	-7	34	-8	-26	3608
Senate-House	0				978	247	-1225			495	2979	-3474

Table 7. The most important named entities around "united nations" item in first 4 months of 2017.

Period 2017	Entities
Jan to Feb	Trump, Flynn, Bannon, White.House, Pence, Abe, Mar.a.Lago, Michael.T.Flynn, Kushner, Stephen.K.Bannon, Russia, America, Japanese, National.Security.Council, Miller, Islam, One.China, McCain,General.McMaster, President George. W. Bush, Sweden, Ninth Circuit, Defense Secretary Jim.Mattis, Taiwan, Palm Beach
Feb to Mar	Ryan, Schumer, Democrats, California, Reid, Justice Department, Donald.Trump, Democratic Party, Senate, liberal, Democratic, Warren, Capitol Hill, Levin, McConnell, Kislyak, DeVos, Republican , North Carolina, Klain, nationalist, Greenblatt, Pennsylvania, Tea Party, EPA
Mar to Apr	Senate, Schumer, Ryan, Democrats, Supreme Court, Reid, California, Donald Trump, Republican, Democratic.Party, Warren, Capitol Hill, McConnell, Levin, Flynn, Kislyak, Texas, North Carolina, Netanyahu, Klain, Merkel, Pennsylvania, Congress, Obama

4.2. NER Sentiment Analysis

In our study, we have collected tweets about North Korea after August 2017 North Korean crisis[1]. We have selected this topic because of its popularity and its footprint. In addition, this topic's lifespan has been expanded to quite long time and this situation make this topic to a really good candidate for a temporal network analysis. We can analyze the data for specific periods, such as weekly, monthly, or any custom selected interval. Here it is worth mentioning that it is not necessary to have intervals equal. We initially considered the first 500 entities based on their frequency. Next, we obtained the temporal co-occurrence between entity results for the designated periods. We have selected weekly analysis for the periods as below:

- Week1 (415 entities): 2017-07-31 - 2017-08-06.
- Week2 (670 entities): 2017-08-07 - 2017-08-13.
- Week3 (472 entities): 2017-08-14 - 2017-08-20.
- Week4 (381 entities): 2017-08-21 - 2017-08-27.
- Week5 (551 entities): 2017-08-28 - 2017-09-03.

We have applied eigenvector centrality to retain the sub network based on the first 100 important nodes.

[1]https://en.wikipedia.org/wiki/2017_in_North_Korea

Each node has the suffx (p: positive or n: negative) to reflect the contribution of the entity in the tweet sentence. The results shown in Figures 15 and 16 report the mapping of nodes for the first and second weeks based on Louvain community detection algorithm (resolution=1). Each community has been represented with a different color. Further, we have summarized the differences between entities in the first two consecutive weeks as reported in Table 8.

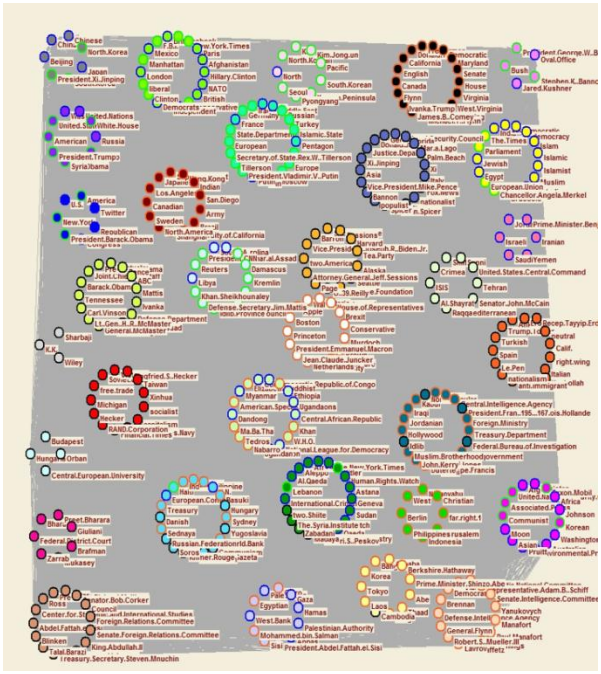

Figure 14. Named entity network created with most popular entities during April to May.

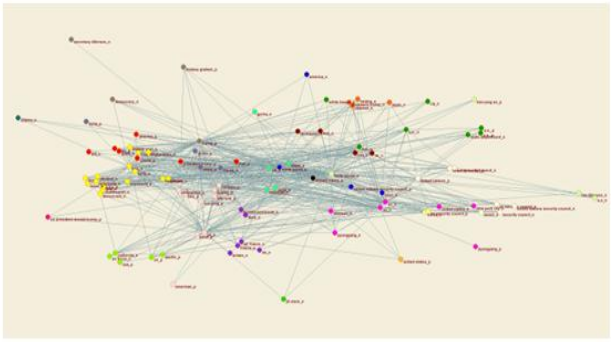

Figure 15. Named entity sentiment analysis network created with all entities during Week 1 time period.

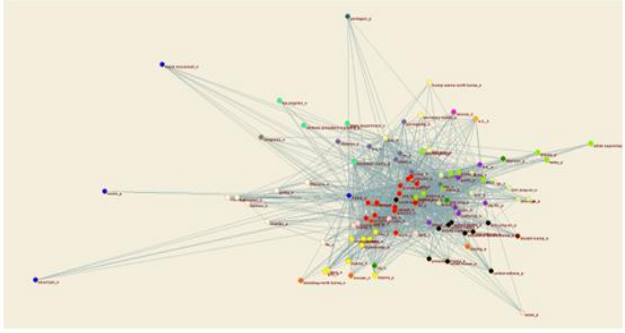

Figure 16. Named entity sentiment analysis network created with all entities during Week 2 time period.

Table 8. List of temporally differing entities in two consecutive weeks.

N	Week(N-1)-WeekN	WeekN-Week(N-1)
2	air force_n, american_p, asean_n, beijing_n, bill_n, britain_n, cuba_n, france_n, iran_p, isis_n, korean peninsula_n, new york city_n, pakistan_n, philippines_p, putin_p, pyongyang_p, rex tillerson_n, syria_p, u.n._p, ukraine_n, un security council_p, un security council_n, united nations security council_p, united nations security council_n, unsc_n, us president donald trump_p, jill stein_n, gorka_p, secretary tillerson_n, u.n. security council_n, lindsey graham_p, security council_p, security council_n, air france_n	american_n, bannon_n, bombing north korea_n, chinese president xi jinping_p, congress_n, dem_n, disney_p, donald trump_n, fbi_n, guam_p, guam_n, jesus_p, kim jong-un_p, kim jong-un_n, kim jung_p, korea_p, liberal_p, mattis_n, mccain_n, mueller_n, nazi_n, nkorea_n, obama_p, ok_p, pentagon_p, president donald trump_p, washington_p, mitch mcconnell_n, moon_p, trump warns north korea_n, white supremacy_p, haley_p, secretary mattis_n, jonathan soble_p
3	bombing north korea_n, chinese president xi jinping_p, congress_n, dem_n, disney_p, fbi_n, jesus_p, kim jong-un_p, kim jong-un_n, korea_p, mccain_n, korea_n, ok_p, pacific_p, president donald trump_p, president trump_n, united nations_p, united states_p, usa_p, washington_p, moon_p, trump warns north korea_n, gorka_n, haley_p, secretary mattis_n, jonathan soble_p	american_p, asia_n, attack north korea_n, britain_n, communism_n, democratic_p, dunford_n, hawaii_n, iran_p, isis_n, israel_n, japan_p, mattis_p, seoul_n, south korea_n, taylor swift_p, ukraine_n, yankees_n, john oliver_n, us-south korea_n, kkk_n, charlottesville_p,charlottesville_n, secretary tillerson_n, steve bannon_p, attack guam_n

5. Conclusions and Future Works

The study described in this paper presented a service-based architecture to handle textual data. Though the study focused on the analysis of online news, it may be easily and smoothly applied to any corpus of textual data. Indeed, regardless how the available data has been captured the process does not change and the discoveries will be equally beneficial. A service-based architecture turned the whole process into platform independent, and hence adds to the flexibility and applicability of the framework. NLP has been used to extract entities and topics regardless of sentiment orientation. The outcome has been enriched by providing the possibility to retrieve statistical summaries for different time intervals. Such summaries enabled the construction of a social network which could be further analyzed for other time intervals.

In the future, new algorithms, libraries and text analysis techniques will be integrated to the proposed pipeline to make more detailed analysis. In addition to them, with using entity co-referencing and finding association rules, we are planning to extract additional layers of information from the text data.

In this study, we focused on a specific date and events to test our service pipeline and analysis tools,

but we are planning to focus and analyse a data set which represents a larger time interval to make more interesting analysis for different topics. Comparing different entity networks for different data sets to find common patterns with using different techniques and discovering additional patterns also be a part of our future works.

Finally, we are planning to create different temporal networks with using different pipelines with different configurations and we are planning to compare results of those different pipelines from different point of views such as accuracy, performance and simplicity to select optimized approach.

References

[1] Agerri R., Bermudez J., and Rigau G., "Multilingual, Efficient and Easy NLP Processing with IXA Pipeline," *in Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, pp. 5-8, 2014.

[2] Baldridge J., "The OpenNLP Project," http://opennlp.apache.org/index.html, Last Visited, 2021.

[3] Batista F. and Figueira A., "The Complementary Nature of Different NLP Toolkits for Named Entity Recognition in Social Media," *in Proceedings of 18th EPIA Conference on Artificial Intelligence*, Porto, pp. 803-814, 2017.

[4] Blei D. and Mcaulidde J., "Supervised Topic Models," *in Proceedings of the 20th International Conference on Neural Information Processing Systems*, Vancouver, pp. 121-128, 2007.

[5] Combe D., Largeron C., Egyed-Zsigmond E., and Géry M., "A Comparative Study of Social Network Analysis Tools," *Web Intelligence and Virtual Enterprises*, 2010.

[6] Dawoud K., Jarada T., Almansoori W., Chen A., Gao S., Alhajj R., and Rokne J., *Handbook of Computational Approaches to Counterterrorism*, Springer Link, 2013.

[7] Feng Y., Abdelli A., Rizzo G., and Troncy R., "Sentinel," https://github.com/D2KLab/sentinel, downloaded, Last Visited, 2021.

[8] Hagen M., Potthast M., Büchner M., and Stein B., "Webis: An Ensemble for Twitter Sentiment Detection," *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, pp. 582-589, 2015.

[9] Han J., Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1-12, 2000.

[10] Hirschberg J. and Manning C., "Advances in Natural Language Processing," *Science*, vol. 349, no. 6245, pp. 261-266, 2015.

[11] Jan-van-Eck N. and Waltman L., "Citation-based Clustering of Publications Using CitNetExplorer and VOSviewer," *Scientometrics*, vol. 111, no. 2, pp. 1053-1070, 2017.

[12] Mahalakshmi G., Vijayan V., and Antony B., "Named Entity Recognition for Automated Test Case Generation," *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 112-120, 2018.

[13] Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., and McClosky D., "The Stanford CoreNLP Natural Language Processing Toolkit," *in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, pp. 55-60, 2014.

[14] Meng Z., Temporal and Semantic Analysis of Richly Typed Social Networks from User-generated Content Sites on the Web, Theses, *University of Nice Sophia Antipolis*, 2016.

[15] Meng Z., Gandon F., Zucker C., and Song G., "Detecting Topics and Overlapping Communities in Question and Answer Sites," *Social Network Analysis and Mining*, vol. 5, no. 1, pp. 1-27, 2015.

[16] Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J., "Distributed Representations of Words and Phrases and their Compositionality," *in Proceedings of the 26th International Conference on Neural Information Processing Systems*, Red Hook, pp. 3111-3119, 2013.

[17] Mrvar A. and Batagelj V., "Analysis and Visualization of Large Networks with Program Package Pajek," *Complex Adaptive Systems Modeling*, vol. 4 no. 6, 2016.

[18] Nadeau D. and Sekine S., "A Survey of Named Entity Recognition and Classification," *Lingvisticæ Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.

[19] Pang B. and Lee L., "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[20] Pinto A., Gonçalo-Oliveira H., and Oliveira Alves-A., "Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text," *in Proceedings of 5th Symposium on Languages, Applications and Technologies*, Dagstuhl, pp. 1-16, 2016.

[21] Röder M., Both A., and Hinneburg A., "Exploring the Space of Topic Coherence Measures," *in Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Shanghai, pp. 399-408, 2015.

[22] Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S., and Tsujii J., "BRAT: A Web-based Tool for NLP-assisted Text Annotation," *in Proceedings of the Demonstrations at the 13th*

Conference of the European Chapter of the Association for Computational Linguistics, Avignon, pp. 102-107, 2012.

[23] Taylor A., Marcus M., and Santorini B., *Treebanks*, Springer Link, 2003.

[24] Wasserman S. and Faust K., *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.

Onur Can Sert received his BSc, MSc, and PhD degrees from the computer engineering department in TOBB University of Economics and Technology. His research interests are big data, natural language processing, machine learning and data mining.

Sibel Tarıyan Özyer received her BSc and MSc from Cankaya University, and PhD degree from Atilim University. She is currently working at R&D Department of Rakun Informatıcs and R and D Inc. Her research interests are social networks, computer networks, internet of things and cloud computing.

Deniz Beştepe received his BSc degree and pursuing his, MSc degree both at the computer engineering department in TOBB University of Economics and Technology. His research interests are web data analysis, machine learning, data mining and medical data analysis.

Tansel Özyer is an associate professor of Computer Engineering at TOBB University of Economics and Technology, Turkey. He completed his PhD in Computer Science, University of Calgary. He received his MSc and BSc from Computer Engineering departments of METU and Bilkent University. Research interests are data mining, machine learning, bioinformatics, XML, mobile databases, and computer vision.