A tweets classifier based on cosine similarity

Carolina Fócil-Arias[1], Jorge Zúñiga[1], Grigori Sidorov[1], Ildar Batyrshin[1], and Alexander Gelbukh[1]

CIC, Instituto Politécnico Nacional (IPN), Mexico City, Mexico
focil.carolina@gmail.com, zujorge@live.com, sidorov@cic.ipn.mx, batyr1@cic.ipn.mx, www.gelbukh.com,

Abstract. The **2017 Microblog Cultural Contextualization** task consists in three challenges: (1) Content Analysis, (2) Microblog search, and (3) TimeLine illustration. This paper describes the use of cosine similarity, which is characterized by the comparison of similarity between two vectors of an inner product space. This research used two approaches: (1) word2vec and (2) Bag-of-Words (BoW) for extracting all relevant tweets to each event related to the four festivals: Charrues, Transmusicales, Avignon and Edinburgh.

Keywords: cosine similarity, natural language processing, Bag-of-Words, word2vec, opinion mining, information retrieval

1 Introduction

Opinion mining is defined as *"the task of classifying texts into categories depending on whether they express positive or negative sentiment, or whether they enclose no emotion at all"* [1].

The growth of social media provides a domain of great interest for several studies related on opinion mining (also known as sentiment analysis) [2] such as opinion analysis related to topics or problems of political preferences [3], opinion about a specific product [4], news [5], and others. According to [6], Twitter is the most popular microblogging network in the world; this microblogging service has more than 300 million users, which represents a competitive advantage for many organizations.

In this project we propose the usage of cosine similarity with two features: Bag-of-Words and word2vec using a dataset of workshop **Microblog Cultural Contextualization 2017** to determine the tweets relevance according to each event from four European festivals.

The remainder of this paper is structured as follows. In section 2, related work on timeline illustration, cosine similarity and opinion mining are presented. Section 3 is focused on showing the materials and methods. Section 4 describes the experimental results where two features: Bag-of-Words and word2vec are used with cosine similarity. Finally, section 5 gives a summary of this work.

2 Related Work

2.1 *TimeLine illustration*

The CLEF 2017 **TimeLine illustration** shared task [7] was dedicated to retrieve all relevant tweets based on festival events. A variety of analysis, such as descriptive (duplicate and web addresses removal, abbreviations, etc.) [8, 9], correspondence and interactive (FactoMineR package) [8] were used.

When looking at the State-of-the-Art on opinion mining for **TimeLine illustration**, we distinguish that Dogra *et al.* [10] used several approaches, such as (i) content-based retrieval, where a query represents a topic for matching in the documents. A common method used in this task is BM25 (Best matching), which is a probabilistic information retrieval function for documents features such as term and document frequencies, and document length [11, 12], (ii) diversification is used when a retrieval function does not take into account the relations among returned documents; they may have relevant and redundant information [13], and (iii) re-ranking for improving the retrieval results through a baseline system [14].

In related research, Murtagh [8] used correspondence analysis to map the dual spaces of days and hashtags to a latent semantic space with social narratives information, and related work on Pierre Bourdieu news.

According to Hoan [15], the model usage according to base knowledge, makes use of an ontology to identify the relation among tweets, festivals and locations. They used a combination of Stanford NER, festival location and user profile. The tweets comparison related to each festival consists of festivals and properties lists with the tweet content.

2.2 *Cosine similariy method*

In this paper, we select a cosine similarity approach for unsupervised learning, and now, we will present some works related to this method with similar objectives.

Shi and Macy [16] compared a standardized Co-incident Radio (SCR) with Jaccard index and cosine similarity. They chose SCR to map sport league studies, music artists, and congress members that can obtain more followers on Twitter.

AL-Smadi *et al.* [17] provided a semantic text similarity approach using text overlap, word alignment and semantic features to identify the paraphrasing in Arabic news tweets. This semantic text similarity is based on Support Vector Regression, which is used for regression analysis. The method achieved good accuracy according to State-of-the-Art.

Tajbakhsh and Bagherzadeh [18] used cosine similarity to find coincidences among tweets. Also, the results were compared with several semantic-based algorithms such as Shortest path, Wu & Palmer, Lin, JiangConrath, Resnik, Lesk, LeacockChodorow, and Hirst-STOnge.

Other related works that use cosine similarity are presented in [19–21].

2.3 *Sentiment analysis approaches*

In this section, we overview several researches based on tweets classification. Looking at particular systems, Dela Rosa, *et al.* [22] presented a study related on clustering and classifying tweets into different categories such as highly relevant, somewhat relevant, not relevant and spam using hash-tags as indicators. The aim of this approach is to find the most representative tweets for some story such as video's Lady Gaga, Obama's help to Japan, disturbing video of Charlie Sheen, and others. The results of this technique performed well based on State-of-the-Art.

Nádia *et al.* [23] besought a combination of classifier ensembles (Random Forest, Suppport Vector Machines, Multinomial Naive Bayes and Logistic Regression) and lexicon for predicting whether a tweet is positive or negative concerning a query term. This ensemble method used two feature representations (Bag-of-Words and hashing) and the results provided an improvement in the State-of-the-Art

Another approach was presented in [24], where a combination of classifier ensembles (Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression) and lexicons were trained for identification of tweets polarity.

In related research, Paltoglou and Thelwall [25] proposed an unsupervised approach based on lexicon-classifier using Bag-of-Words as features. This method estimates the polarity and subjectivity of informal texts on the web, such as tweets, social network and online discussions. The results show that the proposed algorithm presents a trustworthy solution for analyzing feelings of informal communication on internet.

With regard to sentiment analysis of tweets posted on Twitter during a disaster, Venkata *et al.* [26] provided an approach with Bag-of-Words, polarity clues, emoticons, internet acronyms sentistrength and puntuaction as features for identifying and categorizing the sentiments.

As we can see, many researchers have focused on the information that twitter can express, due to the opinions of consumers concerning brands and products [24], political and social events [27], health care [28, 29], higher education [30], microblogging and social network services [31], foresight [32] and others.

In this work, we selected an approach for unsupervised learning called cosine similarity method using two types features: word2vec and Bag-of-Words. Each experiment was evaluated individually.

3 Materials and Methods

3.1 *Dataset*

The dataset for this study came from the workshop **Microblog Cultural Contextualization 2017** [33]. The data consists of festivals and topics collected during July and December 2015. A brief summary of the dataset is shown below.

1. Dataset *"clef microblogs festival"*: This dataset has the tweets related to four festivals, where there are two French Musical festivals, one French theater festival and one Great Britain theater festival. Altogether, this data contains 17.1 GB of information, which is represented by several variables such as:
   - id
   - username
   - date
   - content tweet
   - tweet link
   - microblogging name
2. Dataset *"clef mc2 task3 topics"*: This dataset is a XML file which contains the four festivals. From this set, there are 664 events, which are divided by 61 events that correspond to the *Charrues* festival; 138 to the *Edinburgh* festival, 365 to the *Avignon* festival, and 100 to the *Transmusicales* festival. An example of structure in the XML document is given below.

```
<topics>
    ...
    <topic>
        <id>5</id>
        <title></title>
        <artist>Klangstof</artist>
        <festival>transmusicales</festival>
        <startdate>04/12/16-17:45</startdate>
        <enddate>04/12/16-18:30</enddate>
        <venue>UBU</venue>
    </topic>
    ...
</topics>
```

3.2 *Preprocessing steps*

Tweets gathered from Twitter without a preprocessing stage have noise, confusing words, URLs, stop words and so on. Therefore, this study used the following preprocessing:

1. Consider tweets, which are written in latin alphabet such as English, Spanish, German, Italian, French, and others.
2. Eliminate all retweets.
3. Remove special characters, accents and lingüistic inflections.
4. Convert the dates mm/dd/yy into an only format dd/mm/yy.
5. Serialize and reduce the dataset into 10,000 tweets due to the large amount of tweets.
6. Tokenize all words via NLTK toolkit [34].
7. Replace several white spaces into a white space.

3.3 *Features*

This study uses the following features after preprocessing the tweets.

1. Bag-of-Words. This is a basic representation of words as features. The aim is to compute the word frequencies in a document to find the similarities and differences among the documents [35]. Figure 1 shows Bag-of-Words approach applied in this study.

Document1:
Tweet
I love Anna Calvi #Charrues festival

Document2:
Topic
Anna Calvi, Charrues, 16/07/18

	I	**Love**	**Anna**	**Calvi**	**Charrues**	**Festival**	**16/07/18**
Tweet	1	1	1	1	1	1	0
Topic	0	0	1	1	1	0	1

Fig. 1: An example of Bag-of-Words approach

2. Word2Vec. This is a useful technique used to create word embedding [36]. This means, representing a word as a vector [37]. We use Word2Vec to create an embedding vector for each word in the tweet, compare all vectors and take the maximum value to obtain sentence representation.

I	love	Anna	Calvi
0.23439999	**0.69545599**	0.19999999	0.22222222
0.49999999	0.33339999	0.44566698	**0.78888888**
0.49999999	0.55555555	**0.99999999**	0.11111111
...	...	...	...
0.89999999	0.11111111	0.29999999	0.19999999
0.11999999	0.33333333	0.09999999	**0.77777777**

Generated vector
0.69545599
0.78888888
0.99999999
...
0.89999999
0.77777777

Fig. 2: An example of Bag-of-Words approach

3.4 Cosine similarity

Cosine similarity is a measure that calculates the angle cosine between two vectors [16]. This technique indicates the degree of similarity between documents

which are represented by vectors; when the two vectors are equal then the similarity is high and we obtain a value of 1.

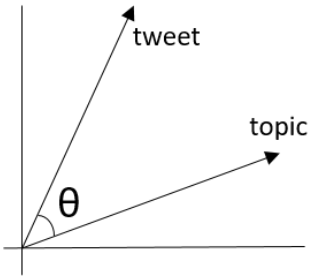

Fig. 3: An example of cosine similarity

In this context, each topic and tweet are represented as vectors, where each vector has the word frequencies (See Figure 1), and then, the cosine formula is applied as follows.

$$cos = \frac{\sum_i^n A_i \times B_i}{\sqrt{\sum_i^n (A_i)^2} \times \sqrt{\sum_i^n (B_i)^2}}, \quad (1)$$

where the topic document is represented by A, and the tweet document is represented by B.

4 Experimental results

In these experiments, we used the format *.res* for representing the results as shown in Table 1.

Table 1: Example of .res format

id_event	id_tweet	rank	score	team_name	run_id
1	634230797378019328	0	0.674039180188	CICLLN	Run1

We conducted two experiments. The first set of experiments consisted in using Bag-of-Words. Based on this approach, Table 2 shows the ten tweets most relevant according to the Charrues event.

In the next stage of experiments, we used word2vec as features to perform the task. The results of this experiment are shown in Table 3.

Table 2: Results of cosine similarity using Bag-of-Words approach

id_event	id_tweet	rank	score	team_name	run_id
1	617440895923810304	0	0.381385035698	CICLLN	Run1
1	598160864143900672	1	0	CICLLN	Run1
1	599058548878921728	2	0	CICLLN	Run1
1	599715191341850624	3	0	CICLLN	Run1
1	609679635245174785	4	0	CICLLN	Run1
1	599996065455198208	5	0	CICLLN	Run1
1	602859447040483328	6	0	CICLLN	Run1
1	616665773579333633	7	0	CICLLN	Run1
1	624511773865984000	8	0	CICLLN	Run1
1	614053444761092096	9	0	CICLLN	Run1

Table 3: Results of cosine similarity using word2vec approach

id_event	id_tweet	rank	score	team_name	run_id
1	624511773865984000	0	0.956182887468	CICLLN	Run2
1	609693683177406464	24	0.903696114115	CICLLN	Run2
1	630692572768432128	32	0.9	CICLLN	Run2
1	599623807641460736	45	0.894427191	CICLLN	Run2
1	631057027259731968	53	0.877058019307	CICLLN	Run2
1	602586360931749889	5	0.956182887468	CICLLN	Run2
1	690208102843576320	59	0.848528137424	CICLLN	Run2
1	614053444761092096	72	0.843274042712	CICLLN	Run2
1	602487688223072256	111	0.822192191644	CICLLN	Run2
1	603514162614915072	353	0.701646415446	CICLLN	Run2
1	603186417154523136	999	0.377123616633	CICLLN	Run2

5 Conclusions and future work

The motivation of this work is focused on retrieval of the most relevant tweets for each of the four festivals. We concentrated on two types of features: Bag-of-Words and word2vec using the cosine similarity approach. Our results show that this technique is capable of detecting the more popular opinion gathered from Twitter based on similarity between tweets and topics.

There is much future work to be done in this study such as the analysis of hashtags, URL, emoticons and tweets characteristics. Furthermore, we will conduct experiments using Bag-of-Words and word2vec features with other measures of structural similarity such as Jaccard index and Phi coefficient.

References

1. Tsirakis, N., Poulopoulos, V., Tsantilas, P., Varlamis, I.: Large scale opinion mining for social, news and blog data. Journal of Systems and Software **127** (2017) 237–248

2. Sun, S., Luo, C., Chen, J.: A review of natural language processing techniques for opinion mining systems. Information Fusion **36** (2017)
3. Golbeck, J., Hansen, D.: A method for computing political preference among twitter followers. Social Networks **36** (2014) 177–184 Special Issue on Political Networks.
4. Fernandes, R., DŚouza, R.: Analysis of product twitter data though opinion mining. In: 2016 IEEE Annual India Conference (INDICON). (2016) 1–5
5. Jumadi, Maylawati, D.S., Subaeki, B., Ridwan, T.: Opinion mining on twitter microblogging using support vector machine: Public opinion about state islamic university of bandung. In: 2016 4th International Conference on Cyber and IT Service Management. (2016) 1–6
6. Oliveira, N., Cortez, p., Areal, N.: The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert System with Application **73** (125-144)
7. Jones, G.J.F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N.: Experimental ir meets multilinguality, multimodality, and interaction, 8th International Conference of the CLEF Association, CLEF 2017, (2017) 11–14
8. Murtagh, F.: Semantic mapping : Towards contextual and trend analysis of behaviours and practices, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016) 1207–1225
9. Chaham, Y.R., Scohy, C.: Tweet data mining: the cultural microblog contextualization data set, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016) 1246–1259
10. Dogra, N., Mulhem, P., Amer, N.O.: Lig at clef 2016 cultural microblog contextualization : Timeline illustration based on microblogs pre-processing of the official tweet corpus content-based matching, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016) 1201–1206
11. Svore, K.M., Burges, C.J.: A machine learning approach for improved bm25 retrieval, Proceeding of the 18th ACM conference on Information and knowledge management (2009) 1811–1814
12. Robertson, S.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval **3** (2010)
13. Zheng, W., Fang, H.: A comparative study of search result diversification methods, Proc. of DDR (2011) 55–62
14. Qi, S., Luo, Y.: Object retrieval with image graph traversal-based re-ranking. Signal Processing: Image Communication **41** (2016)
15. Thi, H., Ngoc, B., Mothe, J.: Building a knowledge base using microblogs : the case of festivals and location-based events, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016) 1226–1237
16. Shi, Y., Macy, M.: Measuring structural similarity in large online networks. Social Science Research **59** (2016) 97–106 Special issue on Big Data in the Social Sciences.
17. AL-Smadi, M., Jaradat, Z., AL-Ayyoub, M., Jararweh, Y.: Paraphrase identification and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features. Information Processing & Management **53** (2017) 640 – 652
18. Tajbakhsh, M.S., Bagherzadeh, J.: Microblogging hash tag recommendation system based on semantic tf-idf: Twitter use case, 4th International Conference on Future Internet of Things and Cloud Workshops (2016) 252–257

19. Pandey, N.: Density based clustering for cricket world cup tweets using cosine similarity and time parameter, 2015 Annual IEEE India Conference (INDICON) (2015) 1–6
20. Kaur, N., Gelowitz, C.M.: A tweet grouping methodology utilizing inter and intra cosine similarity, 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE) (2015) 756–759
21. Prasetyo, V.R., Winarko, E.: Rating of indonesian sinetron based on public opinion in twitter using cosine similarity, 2016 2nd International Conference on Science and Technology-Computer (ICST) (2016) 200–205
22. K.D., R., Shah, R., Lin, B., Gershman, A., Frederkin, R.: Topical clustering of tweets, Proceedings of the ACM SIGIR: SWSM (2011)
23. Da Silva, N.F.F., Coletta, L.F.S., Hruschka, E.R., Hruschka, E.J.: Using unsupervised information to improve semi-supervised tweet sentiment classification. Information Sciences (2016)
24. Da Silva, N.F.F., Hruschka, E.R., Hruschka, E.J.: Tweet sentiment analysis with classifier ensembles. Decision Support Systems (2014)
25. Paltoglou, G., Thelwall, M.: Twitter, myspace, digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology **3** (2012)
26. Neppalli, V.K., Caragea, C., Squicciarini, A., Tapia, A., Stehle, S.: Sentiment analysis during hurricane sandy in emergency response. International Journal of Disaster Risk Reduction (2017)
27. Adedoyin-Olowe, M., Gaber, M.M., Dancausa, C.M., Stahl, F., Gomes, J.B.: A rule dynamics approach to event detection in twitter with its application to sports and politics. Expert Systems with Applications **55** (2016) 351 – 360
28. Kelly, B.S., Redmond, C.E., Nason, G.J., Healy, G.M., Horgan, N.A., Heffernan, E.J.: The use of twitter by radiology journals: An analysis of twitter activity and impactÂ factor. Journal of the American College of Radiology **13** (2016) 1391 – 1396
29. Cardona-Grau, D., Sorokin, I., Leinwand, G., Welliver, C.: Introducing the twitter impact factor: An objective measure of urology's academic impact on twitter. European Urology Focus **2** (2016) 412 – 417
30. Tang, Y., Hew, K.F.: Using twitter for education: Beneficial or simply a waste of time? Computers & Education **106** (2017) 97 – 118
31. Chandra Pandey, A., Singh Rajpoot, D., Saraswat, M.: Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management **53** (2017)
32. Kayser, V., Bierwisch, A.: Using twitter for foresight: An opportunity? Futures **84** (2016) 50 – 63
33. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.y., Sanjuan, E.: Cultural micro-blog contextualization 2016 workshop overview : data and pilot tasks, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016) 1797–1200
34. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. (2009)
35. Kim, H.K., Kim, H., Cho, S.: Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. Neurocomputing (2017) –
36. Ren, Y., Wang, R., Ji, D.: A topic-enhanced word embedding for twitter sentiment classification. Information Sciences **369** (2016) 188 – 198

37. Enríquez, F., Troyano, J.A., López-Solaz, T.: An approach to the use of word embeddings in an opinion classification task. Expert Systems with Applications **66** (2016) 1 – 6