Detection of Sentiment Polarity of Unstructured Multi-Language Text from Social Media

Saad Ahmed, Saman Hina, Raheela Asif
Department of Computer Science
NED University of
Engineering and Technology
Karachi, Pakistan

***Abstract*—In recent years, Twitter has caught the attention of many researchers because of the fact that it is growing very rapidly in terms of number of users and also all the data present as tweets on twitter is public in nature while other social media networks such as Facebook, data is not completely public as users can restrict their post to only users present in their friend list. In this research study, aspect based sentiment analysis (ABSA) was done on the data acquired from social media related to the major cellular network companies of Pakistan (Telenor Pakistan, Mobilink Jazz, Zong, Warid and Ufone). For this research, we have specifically selected all tweets which are not only in English and Roman Urdu but also mixture of above two languages. We have employed natural language processing (NLP) techniques for pre-processing the dataset and machine learning (ML) techniques to detect the sentiments present in the data. The results are interesting and informative specially for policy makers of cellular companies. These companies can utilize this information to increase the performance of their services. In comparison with the state of the art algorithms, the performance of bagging algorithm with this framework on the acquired dataset has produced F Score of 92.25, which is very encouraging outcome of this research work.**

Keywords*—Social media; sentiment analysis; data mining; cellular networks

I. Introduction

As the advancement in science and technology continues, the research plays a vital role in every science and technology related field. This work of research is done on the social media data associated with telecommunication domain. Twitter, a micro blogging website is one of the main stream social media website, which has seen tremendous growth in last few years. In a developing country like Pakistan, common people have now gained access to the Internet and are learning the advantages of social media as a source of information as well as using the same to express their views and ideas about politics, products and services. This makes social media a main source of user generated information which makes it a valuable source of data to perform opinion mining and sentiment analysis of general public.

In the last few years, researchers are working on social media data to extract information and then analyze it using different techniques. Some methods of sentiment analysis have been developed in areas of different domains but still a lot of research needs to be done.

The social media has become a vital part of everyday life where its users can express their ideas, views or comments about any product or service [1]. These views and comments about products and service are very important for companies which are the provider of those products and services. This information from social media can help these companies to refine their strategies for the improvement of their products and services.

Twitter, a micro-blogging real time social media network data is extracted from its website in this research. Twitter generates huge amount of data, this data is extremely valuable for data mining and analyzing sentiments of public. The simplicity of posting tweets in Twitter makes it a suitable data source for real-time sentiment analysis.

Twitter has about 300M+ active users who post about 500M tweets in a single day. This huge data which is generated by users is public and is easily available through APIs (Application Program Interface) to anyone who wants to use this data for analysis. That is why twitter is very popular among research scientists for research purposes. There are several features of twitter such as tweets are maximum of 140 characters, mentions (@) and hashtags (#) which are used by users to refers to any particular event or a company in their tweet. This can be used to collect tweets related to a particular event or company. Tweets have short length, use of local languages and local terms makes it more challenging to analyze and find out the sentiments and possible aspects present in it.

The Twitter is an important source of data acquisition, but it is very complex analyzing its content as large number of the tweets either use slang language or shorten words. Sentence level and word level polarity classification [2] was done using a method based on lexicons, namely, SentiCircles, which builds a dynamic depiction of words in order to determine their suitable semantics. Here, semantics refers to the co-occurrence patterns from each word in the dataset. A different method is feature engineering [3] which produces a result of seven dimensions. This feature engineering method was used to analyze aspects: frequency, affinity, valence, shifter, feature sentiment scoring and categorization. Different type of representations can be utilize, based on dictionaries and lexical aspects of sentences [4], word embedding [5], word and character n-gram [6] among others.

The extraction and classification of user opinion on the diverse topics is known as sentiment analysis which is also referred as opinion mining. Mostly, two forms of methods are used for sentiment Analysis, which are either based on machine learning or based on vocabulary. The machine learn-

ing method has two approaches, supervised learning and unsupervised learning. Supervised learning involves data which is labeled to train algorithms [7], whereas unsupervised learning does not need data to be labeled [8]. The mixture of labeled and unlabeled data makes semi-supervised learning [9].

We proposed a hybrid sentiment analysis framework. This method comprises of a customized dictionary of Roman Urdu words which are commonly used by social media users in Pakistan to express their views and share their comments on twitter. This has helped us extract more information from tweets and use this additional information [10] to detect the sentiments of the people. In addition to this, the proposed framework also includes the use of SENTIWORD dictionary which provides weights for each English word which appears in the tweet. By using these weights we were able to calculate more realistic sentiment polarity which improves overall performance of this framework.

Since the beginning of 21st century, Sentiment Analysis (SA) has become one of the main area of research in natural language processing (NLP) [11]. It is a complex problem with many exciting sub-problems, which includes sentence-level sentiment classification, which is the case in tweets. Research scientists have documented that different type of sentence require different handling for SA. Sentence can be of different types, which includes subjective sentences, comparative sentences, negation sentences, conditional sentences, target-dependent sentences and sarcastic sentences. The tweets extracted from twitter could carry all these types of sentences which present a more complex problem during analysis.

Aspects [12] are basically features; these are selected by using Information gain method. The sentiment of the feature is calculated by using the neighboring words of the aspect. These are acquired through N-gram methods. To calculate the effectiveness of this hybrid method, we obtained a corpus from Twitter, we collected data for the duration of six months from 15 Dec 2016 to 14 June 2017 and 2703 tweets were extracted which were then manually classify as positive, negative or neutral.

Our experimental results confirms that the good result was obtained through the N-gram around method [13] along with the use of customize Roman Urdu dictionary. In addition to this, documents such as customer reviews may contain fine-grained emotion for different features (e.g. a product or service) that are mentioned in the document. This information can be very valuable for understanding customers' opinion about a certain service or product on twitter.

Twitter has seen rapid growth in the last few years, where its registered users can post tweets related to events in real time [14]. Users of social media tend to tweet using highly unstructured language with many typographical errors and use local languages as well as slang words in tweets. A significant amount of tools and setup is required to work on social media data due to its speedy growth and to the difficulty of processing its data by using standard relational SQL databases [15].

Today, social media users share their views and opinions on internet, increasing the volume of information each day. Social networks like Twitter and Facebook sites are most popular. Facebook [16] reaches its 1 billion users in October 2012, while twitter had more than 500 million users on February 2012 [17], [18] and currently it has more than 690 million registered users and on average twitter recorded 9100 tweets/second.

According to www.Twitter.com, the most discussed topics [19] in 2016 are as follows:

1) Rio2016.
2) Election2016.
3) PokemonGo.
4) Euro2016.
5) Oscars.
6) Brexit.
7) BlackLiveMatter.
8) Trump.
9) RIP.
10) GameofThrones.

On 24 October 2015, more than 41m tweets related to "#ALDubEBTamangPanahon" of AlDub [20] were sent during special concert of the Kalyeserye segment of the show Eat Bulaga entitled Eat Bulaga: Sa TamangPanahon held at the Philippine Arena, the world's largest indoor arena in Bulacan, Philippines. This performance was attended by more than 55000 people. This was the most discussed topic on twitter.

The top most sports game ever discussed on twitter with over 35.6 million tweets was the 2014 FIFA World Cup semi final held on July 8, 2014 between Brazil and Germany.

Twitter is a very popular social media site for data mining research where there is significant amount of data available containing all type of information. The information regarding followers, followed, tweets, and posts can be used in Recommender system [21] and can also be used to mine valuable information like public mood, trends and sentiments.

ABSA emerges as excellent technique which enables us to find the best solution. Recently ABSA based on social network data is gaining importance in the field of data mining. The result of this proposed framework will reflect the mood of the general public.

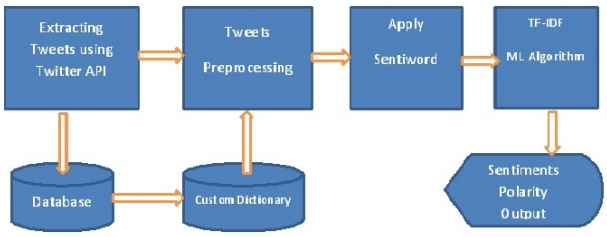

Fig. 1. Proposed system to detect sentiment polarity.

II. Methodology

The main focus of this research work was to develop a hybrid framework as shown in Fig. 1, which employs data mining and machine learning techniques using data from social media network and obtain results which are valuable for Pakistani cellular companies. Main steps involved in this framework are using data crawler to mine twitter website using twitter API. Unstructured data (tweets) is converted into structured data

Algorithm : Sentiment Analysis

```
while Next document do
  for each word in document do
    Remove urls, stop words, numbers, special characters
    if misspelled(word) then
     Replace word with suggested correct word
     end
   end
     Now calculate Sentiment score using SENTIWORD
     dictionary
     Display results and write it on output file
End
```

Fig. 2. Algorithm of sentiment analysis.

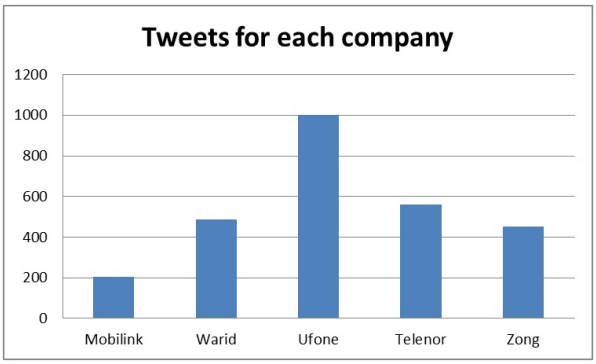

Fig. 3. No. of tweets extracted for each cellular company.

TABLE I. ANALYSIS OF TWEETS COLLECTED

NAME	Total	Negative	Positive	Neutral
Mobilink	204	34	120	50
Warid	487	35	149	303
Ufone	1000	84	764	152
Telenor	561	50	364	147
Zong	451	21	338	92

and is stored in a database. This data includes some irrelevant information which was cleaned by applying preprocessing steps on this data using Natural Language Processing (NLP) techniques and finally we use learning classifiers to find the sentiments of the tweets. We have used R programing language to perform statistical calculation on the dataset. This R platform has provided us all the tools, required for this research work. Algorithm of proposed sentiment analysis framework is shown in Fig. 2.

The algorithm uses SENTIWORDNET [23] for assigning weighted scores to determine the polarity of analyzed tweet.We then apply Machine learning Algorithm and use Term Frequency Inverse Document Frequency (TF-IDF) technique [18] to obtain the aspects and there weights. To achieve constant processing time the Twitter data corpus is divided into parts of equal size in the testing process. The block diagram of the framework is shown in Fig. 1. The process is discussed in depth in the following sections:

A. Data Preprocessing

The dataset which is downloaded from social website Twitter has 2703 tweets which is from 15 Dec 2016 to 14 June 2017 by using the API (application Program Interface) and the Data Crawler. Then on this data, preprocessing was done to clean data by removing the garbage like website address and links to images, which are of no use in sentiment analysis research project.

Total tweets collected were 2703 in dataset out of which 1000 tweets belongs to Ufone, 451 tweets belong to Zong, 561 belong to Telenor, 204 belongs to Mobilink and 487 tweets belongs to Warid as shown in Table I. These tweets were extracted from Timelines of official twitter account of these companies. This is depicted in Fig. 3.

The collected corpus has very detailed information about each tweet, it has 16 fields in it. We retain only the useful and essential fields of data and store it in the .csv files, the fields includes User ID, User Screen Name, Reference, Tweet ID, Date and Time and importantly Text field is stored in database and remaining data fields were filtered out. The Dataset is now structured and is in organized form to be tested.

B. Text Preparation

Tweets (documents) are now parsed into a data corpus for text analysis. Text field of the tweet is considered. The text present in text field is prepared by cleaning for further analysis.

During text preparation, the numbers, URLs and links to images, videos and websites are removed from tweets as they do not serve any purpose. Stopwords such as “but”, “shall”, “by”, etc. are words which have no analytical importance but are commonly used, so these stopwords are removed from the text. After this Stemming process is done to reduce inflected words to their root form which makes system analyze words better, for this purpose suffix dropping algorithms are used in this step. Punctuation marks and whitespaces are removed as they also serve no purpose in sentiment analysis. Lemmatisation algorithms are applied finally to complete the data cleaning process.

In this preprocessing step of the tweets, a large level of noise is removed by using tokenizing which is a process of splitting text into a set of individual terms or tokens. Each tweet is tokenized into a sequence of terms. In NLP, the most commonly used words in a document are referred to as ‘stopwords’. All the tweets are checked against a standard stopword list to remove terms which carry little information. The token starting with ‘@’ (i.e. a reply or mention) will also be removed from the tweets in the filtration process. At the end of this process, each tweet is divided into a set of aspects which are in the vector space model.

C. TF-IDF Technique

In this research work we have utilized TF-IDF technique [14]. We have applied this technique to filter out tweets which have minimal or no information which helps in our analysis during this research.

This technique makes a sparse matrix (a matrix in which most elements are zero), this indicates that how many parameters are un-informative in the dataset. So we reduce sparsity by

removing the terms that occur very regularly. This has shown to have the effect of reducing over fitting and improving the analytical capability of our system.

In this research we choose to set the sparsity to a maximum of 75% which has provide us improved results without bias to the context and perspective of sentiment of public. Tweets with no information to predict the sentiments were also filtered out to improve the performance of the ML algorithm.

Normally data mining frameworks use clustering methods, which groups similar items in a one subset. These subsets of items are called clusters. Different types of techniques such as Nearest Neighbor (NN) and K-mean [22] can perform clustering. The clusters which are created by these algorithms are used as polarities of the opinion or sentiments of people.

III. Experimental Results

The dataset is divided into two same size parts, i.e. training and testing data. The presented system was tested using different ML algorithms with TF-IDF, using the dataset acquired from the twitter.com. The other major feature of this work is that we have develop a list of words which are commonly used by users in their tweets which are not English words but are words written in Roman Urdu. With the help of this customize dictionary of Roman Urdu words, we were able to better understand the context of tweets written using words both from English and Roman Urdu in a single tweet and this results in better detection of sentiments at sentence level and increases the performance of this proposed system.

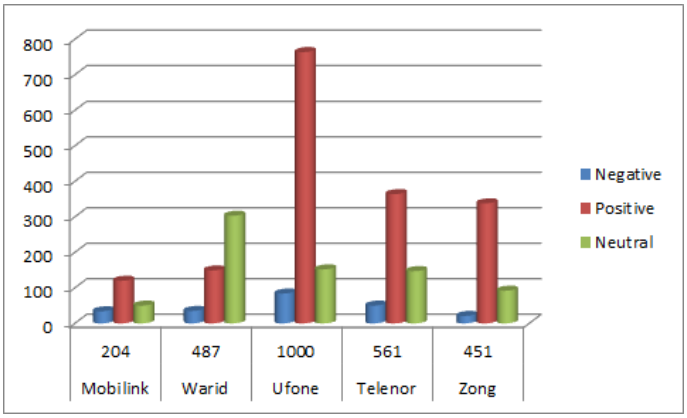

Fig. 4. Results of analysis showing polarity of tweets for each company.

The results obtained are illustrated in Fig. 4, which shows that the positive aspects in the tweets dataset are in abundance as compared to the negative or neutral with very few sarcastic tweets. This identifies that the telecommunication companies are giving better service to their customers in Pakistan. Ufone has the most number of positive sentiments towards their services which is in line with this common perception that Ufone is the most popular cellular network company in Pakistan.

Fig. 5 shows the performance of presented hybrid system to detect sentiment polarity. The services of cellular networks in Pakistan are becoming reliable and dependable, this statement was validated during this research work as the number of tweets predicted which carry positive sentiments are far more than the other sentiment combined.

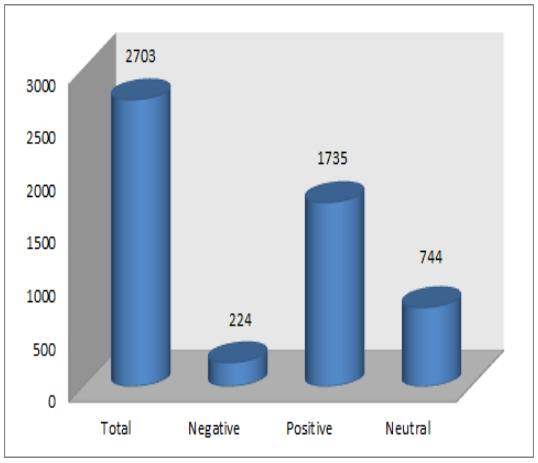

Fig. 5. Polarity of sentiments for all cellular companies.

TABLE II. Performance of Framework

ALGORITHM	PRECISION	RECALL	F SCORE
Boosting	0.9601	0.8925	0.9148
SVM	0.6811	0.5975	0.6311
Bagging	0.9651	0.8975	0.9225
Forest	0.4025	0.3799	0.3875
Tree	0.7001	0.6454	0.6601
Maxen	0.4755	0.6601	0.4925
Naive Bayes	0.8160	0.8150	0.8240

Predicted sentiments of all the tweets used in this research work are illustrated in Fig. 6. The closer look at these predicted polarity of tweets gives us the information that over the period of time during which tweets were collected there is an increase in positive sentiments towards cellular companies in general which is the indicator that services in telecommunication sector is improving very rapidly and customers are getting state of the art technology which is affordable.

The Precision, Recall and F Scores of algorithms which were used to compare the performance are shown in Table II.

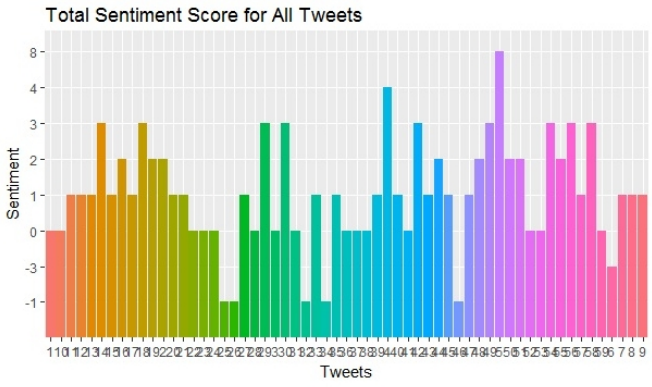

Fig. 6. Sentiments of tweets predicted.

The performance of bagging and boosting algorithm with this framework was superior as compared to other well-known algorithms, as depicted in Fig. 7. The Precision of Nave bayes is more than 81% while SVM and tree algorithms achieved above 60% precision and Forest algorithm achieved only 40%

precision on this dataset.

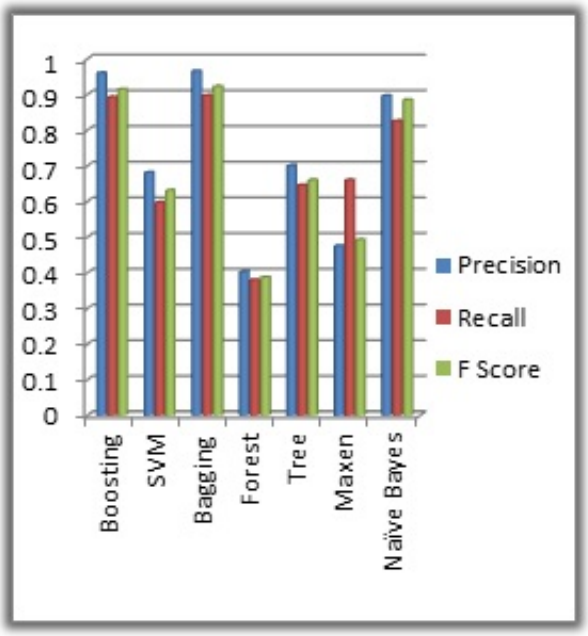

Fig. 7. Comparison of performance of proposed framework with different ML algorithms.

IV. Conclusion

Internet has open doors of information and social networks on internet have provided an important source of data regarding users and their sentiments towards a particular product, service or event. This is especially valuable in giving in depth knowledge about the current developments and attitudes of people who are using Internet. In this paper we have presented and applied a hybrid system of sentiment analysis to analyze tweets from twitter. This hybrid system has been weighed using twitter data related to cellular companies of Pakistan.It will be further evaluated by increasing data set to provide efficient/faster processing on big data.

We have collected real data from social media network Twitter which is related to cellular companies of Pakistan by using the Twitter streaming API. This API also provides detailed meta data for the tweets.

Data preprocessing has been done to improve the accuracy of this hybrid system. The addition of customize Roman Urdu dictionary has added a new dimension to this research work and produced motivating results.We were able to correctly detect the sentiment polarity of the tweets used in this research and these results were confirmed by human annotation of the same data. The framework which we have proposed uses TF-IDF technique with Bagging machine learning algorithm and is identifying the sentiments faster with improved F-scores, this proposed framework has also produced better results with boosting ML algorithm. We have also evaluated the accuracy of this system by comparing methods used by other researches in this area of research and found the performance of this framework is comparable with the other state of the art algorithms.

References

[1] Ravi, Kumar, and Vadlamani Ravi. “A survey on opinion mining and sentiment analysis: tasks, approaches and applications.” Knowledge-Based Systems 89 (2015): 14-46.

[2] Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. Information Processing and Management, 52(1):5 - 19

[3] Ghiassi, Manoochehr, David Zimbra, and Sean Lee. 2016. Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for articial neural networks. Journal of Management Information Systems, 33(4):1034-1058.

[4] Murillo, Edgar Casasola and Gabriela Marin Raventos. 2016. Evaluacion de modelos de representacion del texto con vectores de dimension reducida para analisis de sentimiento. In TASS@ SEPLN, pages 23 - 28.

[5] Quiros, Antonio, Isabel Segura-Bedmar, and Paloma Martnez. 2016. Labda at the 2016 tass challenge task: Using word embeddings for the sentiment analysis task. In TASS@ SEPLN, pages 29-33.

[6] Ceron-Guzman, Jhon Adrian and Santiago de Cali. 2016. Jacerong at tass 2016: An ensemble classifier for sentiment analysis of Spanish tweets at global level. In TASS@ SEPLN, pages 35-39

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP, pages 7986, 2002.

[8] Peter D Turney. Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), (July):417424, 2002.

[9] Andrew B Xiaojin. Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, pages 1130, 2009.

[10] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. “A comparative study of TF* IDF, LSI and multi-words for text classification.” Expert Systems with Applications 38.3 (2011): 2758-2765.

[11] Liu, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.

[12] Salas-Zrate, Mara del Pilar, et al. “Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach.” Computational and mathematical methods in medicine 2017 (2017).

[13] Cavnar, William B., and John M. Trenkle. “N-gram-based text categorization.” Ann Arbor MI 48113.2 (1994): 161-175.

[14] J. Lin and D. Ryaboy, “Scaling big data mining infrastructure: the twitter experience,” SIGKDD Explor.Newsl., vol. 14, pp. 6-19, apr, 2013.

[15] G. Mishne, J. Dalton, Z. Li, A. Sharma and J. Lin, “Fast data in the era of big data: Twitter’s real-time related query suggestion architecture,” in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, USA, 2013.

[16] Ashlee Vance (October 04, 2012) Facebook: The Making of 1 Billion Users. [Online]. Available: http://www.businessweek.com/articles/2012-10-04/facebook-the-making-of-1-billion-users

[17] Lauren Dugan (February 21, 2012) News, Statistics: Twitter to Surpass 500 Million Registered Users on Wednesday. (Online). Available : http://www.mediabistro.com/alltwitter/500-million-registered-usersb18842

[18] Twitter Inc. (2011) Year in Review: Tweets per second. [Online]. Available: http://yearinreview.twitter.com/en/tps.html

[19] https://blog.twitter.com/official/enus/ a/2016/thishappened-in-2016.html

[20] “Fans in the Philippines and around the world sent 41M Tweets mentioning #ALDubEBTamangPanahon”. Twitter Data Verified Account. October 27, 2015. Retrieved October 30, 2016.

[21] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey Universidad Politcnica de Madrid, Ctra. De Valencia, Km. 7, 28031 Madrid, Spain

[22] Hartigan, John A., and Manchek A. Wong. “Algorithm AS 136: A k-means clustering algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[23] sentiwordnet.isti.cnr.it/