*I.J. Information Technology and Computer Science,* **2019, 12, 19-29**
Published Online December 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2019.12.02

# Sentiment Analysis of RSS Feeds on Sports News – A Case Study

**Khalid Mahboob**
Dept. of Software Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan
E-mail: nedian07@gmail.com

**Fayyaz Ali**
Dept. of Computer Science, Sir Syed University of Engineering & Technology, Karachi, Pakistan
E-mail: fayyaz54@gmail.com

**Hafsa Nizami**
Dept. of Software Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan
E-mail: hafsanizami@yahoo.com

Received: 19 August 2019; Accepted: 14 September 2019; Published: 08 December 2019

***Abstract*—**With the advent of online social media, such as articles, websites, blogs, messages, posts, news channels, and by and large web content has drastically changed the way individuals take a glimpse at different things around them. Today, it's an everyday practice for some individuals to read the news on the web. Sentiment analysis (also called opinion mining) alludes to the utilization of natural language processing, content investigation, and computational linguistics to distinguish and separate subjective data in source materials. Sentiment analysis is broadly applied to online reviews, news feeds and social networking for a wide variety of applications, ranging from marketing to client services. Sentiment analysis emphasizes on the classification of textual data into positive, negative and neutral categories. This research is an endeavor to the case study that calculates news polarity or emotions on different sports feeds which may influence changes in sports news development patterns. The interest of this approach is to generate various text analytics that computes feelings from all pertinent ongoing sports news accessible out in the public domain. The significance and application value of sentiment analysis of RSS feeds in this study is to distinguish between positive feeds and negative feeds on sports that could affect readers or users minds in order to improve RSS feeds messaging broadcast among folks. The methodology utilizes the sentiment analysis techniques using two different online open-source sentiment analysis tools in Rich Site Summary (RSS) news feeds that have an influence on sports-related broadcast esteems.

***Index Terms*—**Sentiment Analysis, RSS feeds, Sports News, Polarity, Social Media.

## I. Introduction

The world has become a global village in the current era with the advent of the internet and social networking. Web-based life is any type of online distribution or nearness that permits end users to take part in multi-directional discussions in or around the substance on different websites like posting status, commenting on different blogs, threads or posts, etc. Social media and social networking center on two-way connections, between the webpage (and the individual running the website) and the general population perusing or utilizing it. Two-way collaborations imply that comments are permitted in that blog and there could be associations between the writer and their comments[1, 25]

Since the inception of internet technology and social networking around the 1960s, it has evolved with the discussion or information sharing forum to share opinions and suggestions among different users. Social networking is significant nearly in every field from education to business, to facilitate students as well as teachers/tutors to collaborate with each other using social networks that can enable the students to pick up chances to cover learning in a more extensively and seek interest. A standout amongst the most essential thing to call attention to the effect of informal organizations in the business field is, social media advertising and marketing[13, 25].

By acting social networking communities as stepping stones, social media showcasing can pick up a considerable measure of benefits including relationship building, brand building, publicity, promotions, etc. So, it tends to be inferred that social networking proposes a number of opportunities for entrepreneurs, independent ventures, fair size organizations, and extensive partnerships to fabricate their brands furthermore, business [1, 25].

A user can find out optimal decisions and clutching different types of updated information through online social networking need some information feeds to regularly update a web user with the updated summary of web contents[24]. The aim of studying and analyzing one of the mechanisms of social information is RSS feeds on different websites referring to simply a feed by which a reader may subscribe into site content, for example, a blog or news website [5, 24]. At first, RSS feeds were planned for news headlines. The usage has extended to incorporate discussion features, new sale postings, refreshed postings of houses available to be purchased, and various different employments [7]. There is no need for a web user to spend time searching down the content that is needed and keeping the user updated on updating information occasionally.

Sentiment analysis (once in a while called as an emotion or sentiment evaluation) refers to the practice of natural language rework, textual counter-revolution, computational syntax, and biometrics to systematically understand, extract, quantify, and have a look at operative states and skewed statistics [1]. Sentiment analysis is broadly carried out to opinion of the client substances including critiques and anticipation's, online and social media, and healthcare for programs that vary from advertising and promotion to consumer service to clinical treatment to evaluation of comments or news feeds. Commonly stated, sentiment evaluation desires to determine the mindset of a speaker, author, or different individuals with respect to the overall contextual polarity or emotional response towards document, event, or interaction. The opinion can be a decision or assessment, or the effective expressive conversation [2].

RSS (`Really Simple Syndication' or `Rich Site Summary') was developed by Dan Libby and Ramanathan V. Guha on Netscape basically regard as a group of web feed formats for the purpose of syndicating content from online blog entries, news headlines or web pages to provide variety of information to the reader or user [12]. RSS is an XML-based format for distributing and aggregating web content (such as sports, business, politics, weather, etc. headlines). RSS feeds empower publishers to syndicate a wide variety of data automatically for every individual [6]. RSS is essentially an XML record that outlines data items and links to the data sources [12]. Extensible Markup Language (XML) is basically describing data. The XML standard is typically a way to create data formats flexibly and share structured information electronically through the public internet, as well as by means of corporate networks. XML code, a formal recommendation is similar to Hypertext Markup Language (HTML). Both XML and HTML contain markup symbols to describe page or file content [15]. RSS feeds likewise advantage different users who want to get convenient updates regularly from desired online websites or to aggregate information from numerous websites [6]. RSS document (called "channel", "web feed", or "feed") incorporates full or summarized content, and metadata, such as publishing date and author's name [6].

The ubiquitous evolution in the prevalence and adoption of RSS can be evidently found within the feed add-ons in all major web browsers, and the manifestation of various online aggregators and readers [5]. There are two different versions of RSS e.g. RDF (RSS 1.*) and RSS 2.* have been evolved. The first line of RSS feeds contains the XML tags. The next line shows three RSS feed tag which declares the identification of documents in RSS feeds. It contains the channel tag which shares the whole information of an RSS feeds. The channel tag contains three child tags namely, Title, Link, Description. A feed is organized of a <channel> component which contains metadata about the channel and its substance, i.e. title and description [5]. Inside the <channel> are various <item> components, each of which comprises of the components, some of them are contained in Table 1.

The title tag contains the title of the RSS feeds, the link tag contains the link of the information, and lastly, the description tag contains the short summary of the topic. Thus, in computational processing, a news aggregator, likewise named a channel aggregator, feed per-user, news per-user, RSS per-user or basically aggregator is client programming or a web application which aggregates syndicated web stuff, for example, online daily papers, sites, web recordings, and video web journals (vlogs) in one area for simple survey [16].

The RSS feeds enable a user to stay track of the many completely different websites during a single news collector. The news collector can examine the RSS feed for brand new content, permitting the subject to be automatically delivered from one web site to another web site or from a web site to reader or user [3]. Positive news has been dominated and obtaining more attention. The positivism closes the nice news has been drastically reduced by the amount of unhealthy news. The target of this research is to supply a platform for serving excellent news and build positive surroundings [4]. Basically, RSS feeds collect new content from the web sites visited most often and distributes it to desktop, browser, or e-mail account. Through RSS feeds user can monitor news, blogs, and content for any personal and professional interest. With RSS, Web content comes to use as it is newly published on the Web rather than user needing to go looking for it [5].

Sentiment analysis of online user-generated content is very important for several social media analytics tasks. A great deal of labor has been allotted for extracting individuals sentiments from matter information. Researchers have mostly practiced sentiment analysis to establish a system to predict political elections, live economic indicators, and so on. Although social media is supply most up-to-date info, it cannot be trustworthy because it consists of many aspects generated by totally different people. This research is tending to the area unit proposing a lexicon-based approach of sentiment analysis for sports RSS feeds. The lexicon-based approach consists of aggregating sentiments from each social media and news feeds. Once extracting sentiments from each approach, they're then clustered for analysis [6].

Table 1. Components of <item> RSS feed

| Component | Description |
|---|---|
| <title> | The heading of the channel |
| <description> | An explanation of the item which can be either text or HTML |
| <link> | A hyperlink to the full content on the beginning website |
| <guid> | A unique identifier, often the hyperlink of the full content |
| <pubDate> | The item's publication date for the content of the feed |
| <category> | The category specify item by the publisher. |

## II. Literature Survey

In recent years, significant endeavors have been put into practice that can anticipate the future pattern of particular analyses of the subjectivity on social issues in the form of news, posts, blogs, or feeds, etc.

J. A. Morente-Molinera et al. [1] presented a novel method that allows extracting useful information from Twitter about the debates that are carried out in social networks. The method focused on a group decision making approaches scheme. The sentiment analysis methodology is additionally utilized to change over the normal texts that social network members use to give conclusions into numerical qualities that the framework can deduce. To gauge the dependability of the produced comes about accord measures for various perspectives.

Yashodhara Haribhakta et al. [4] discussed to provide a stage for serving uplifting news and make a positive situation. In this concern, an algorithm was proposed for the classification of different News articles. This incorporates tool of data aggregator and handling engine at the server side as a Sentiment classifier and a stage for a client where positive news being served to peruse. SVM algorithm has been used for building the model as a classifier. Total 480 news articles have been analyzed out of which 265 were positive and 215 were negative.

Martin O'Shea et al. [5] presented two distinct technologies to employ the semi-structured nature of RSS to enable clients to mine data specifically from crude RSS feeds. Two technologies with their implementation experimentally were depicted to mine few RSS feeds and in the direction of visualizing the data mined. An application visualRSS (vRSS), have selected to give a platform to mining and visualizing textual data initially from RSS inside a social information system to later incorporate numeric values and other data sources but further, it can be extended up to include other alternatives like Atom in RSS.

Kalyani D. Gaikwad et al. [6] surveyed a generic overview of opinion mining and sentiment analyses techniques. The primary objective of opinion mining is sentiment classification i.e. to arrange the sentiment into positive or negative classes. Further, RSS utilizes a group of standard online web feed arrangements to broadcast the recurrently updated information: blog passages, news features. Various sentiment analyses methods and their uses have also been discussed such as Naïve Bays Classifier, Support Vector Machine (SVM), Multilayer Perceptron, Clustering.

SV.Shri Bharathi et al. [8] attempted to present a predictive model that can predict news polarity which may influence changes in stock price development patterns. The sentiment analysis approach used is Really Simple Syndication (RSS) news feeds that can have an impact on stock market prices. In this experimental study, the algorithm is used to predict stock price variances, regardless of whether up or down.

Alessia D’Andrea et al. [9] gave an overview of the different sentiment classification approaches and tools to perform sentiment analysis. The sentiment classification approaches can be classified into machine learning, lexicon-based and hybrid approach correspondingly. The most commonly used sentiment tools for detecting the feelings polarity like Emoticons, LIWC, SentiStrengh, Senti WordNet, SenticNet, Happiness Index, AFINN, PANAS-t, Sentiment140, NRC, EWGA, and FRN have discoursed that can be applied in different fields.

R. Kent Wills [10] proposed the making of a production-grade classifier for a website for sentiment analysis. Moreover, the polarity of organized and unorganized information, RSS and Twitter feed, ought to give in-locate in the matter of how one might be operated to support the other. Naive Bayes and Decision Trees methods have been used.

Amruta S. Dulange et al. [11] developed a crawler to peruse information from RSS feeds of different blogs in a websites and save them locally in order to index the user blogs for easy seeking and information extraction by applying certain data mining technique to examine user/customer negative or positive perception using multiple blog sources.

Juana María Ruiz-Martínez et al. [12] proposed an algorithm based methodology that consolidates several gazetteer records and leverages current financial news that acquired from different RSS feeds to annotate positive or negative markers. The result of the procedure is an arrangement of news organized by their level of positivity and negativity. The open-source software GATE was used to carry out a sentiment and semantic annotation using gazetteers lists.

Shri Bharathi et al. [13] attempted designing and implementing a predictive system to manage stock market investment using the combination of sensex points and Really Simple Syndication (RSS) feeds for efficient prediction. For the experimental study, the stock market prices and RSS news feeds are gathered for the organization ARBK from Amman Stock Exchange (ASE). The sentiment polarity associated with the news sentences was calculated for stock news prediction, to determine whether they were positive, negative or neutral.

Emma Haddi et al. [14] examined the sentiment of online movie reviews with the combination of different text pre-processing approaches in order to attain noise reduction in the text using chi-squared method to evacuate insignificant features not influencing its orientation. Further, data transformation and filtering may significantly enhance the performance of the classifier.

Like the works or concepts presented in [5, 6, 8, 11, 12, 13], the idea stated in this current study is to use a centralized web-based system so that all the information of RSS feeds will be available through web-based system access to the end-users. The sentiment analysis on RSS feeds (a case study on sports news) can be performed using two different sentiment analysis tools namely voyant-tools and LIWC (Linguistic Inquiry and Word Count) [9] to determine through different sentiment experiments in order to computationally identify and classify the user opinions impact from the text of RSS feeds on sports news, in particular to determine either the reader's would perceive the specific subject in feeds is positive or negative.

## III. Data Description and Methodology

There are generally two main approaches to sentiment analysis: the one is the lexicon-based approach where we firstly split some text into smaller tokens from the words phrases or whole sentences. This process is simply called the tokenization where we count each word and next we look up the subjectivity of each word from an existing lexicon which is a database of emotional values for words prerecorded by the data dictionary. The overall subjectivity in a text can be analyzed by the lexicon corpus-based approach. To carry out this study, the corpus-based on RSS feeds description containing text on different popular categories of sports such as badminton, cricket, football, hockey and tennis (particularly famous in Pakistan) have been obtained to apply sentiment analysis approach to label RSS feeds into positive emotions and negative emotions respectively.

Ontologies are necessary to establish the standard learning and knowledge representation instrument for the semantic web. It automates information processing to enable the use of semantic reasons to gather new information [12]. In order to collect RSS feeds data, a web-based system was developed to obtain feeds from a remote web server (also accessible through mobile) for users where the different updated news feeds on different sports through RSS feed news aggregator have been automatically generated are collected and extracted on data grid view that has been processed through a remote web server. The Fig. 1 describes the overall system for generating, collecting and processing various news feeds on particularly five sport categories namely: badminton, cricket, football, hockey, and tennis to perform sentiment analysis to differentiate the impact of RSS feeds on a reader/user.

The proposed system consists of a Website and an Android Application which is divided into different modules. The administrator has the permissions to access all the data which is available in the system whether confidential or static. There will be an RSS module where the users are able to login with their id and passwords provided them by a system administrator.

The corpus was built on different RSS feeds of five different sports such as badminton, cricket, football, hockey and tennis that contain at least a hundred news feeds for each sports category. The text mining was performed to deduce quality information from the text of RSS feeds. Only descriptions of RSS feeds (which contain sentences and phrases) have been examined for sentiment analysis to conclude various text analytics through statistical pattern learning. The sports RSS feeds data have been collected from different online news websites like a tribune.pk, samaa.tv, geo.tv, dawn.com, thenews.com.pk, etc. and classified into five sports groups [18- 22].

There are various sentiment analysis tools available to perform text analytics, finding hidden polarities or feelings, finding presence and frequency of terms, information, and rules of parts of speech (POS), pros and cons in a text, and so on [9]. For this study, two online open-source sentiment analysis tools have been used to perform various text analytics by extracting and processing sports RSS feeds description. One of which is a voyant-tools, a web-based text reading, and analysis platform use to learn the scalable computer-aided analysis functions [17]. The corpus can easily be created, modified, and embedded on voyant-tools. The tool language interface can support multiple languages for the analysis of documents or text. There are several tools incorporated in an environment for multiple purpose text mining and analytics like bubblelines, cirrus, correlations, scatterplot, word tree, terms radio, textualarc, streamgraph, etc. [17]. Different groups of sports RSS feeds are integrated into a text corpus to apply certain methodologies of text mining using voyant-tools.

Another online web-based open-source tool employed was LIWC (Linguistic Inquiry and Word Count) is the best quality level in computerized text analysis. It can figure out how the words use in ordinary dialect uncovers the emotions, feelings, behavior, motivations, and thoughts. Based on scientific research, LIWC is more precise, less demanding to utilize, and gives a more extensive scope of social and psychological awareness of knowledge [23]. The tool permits investigating positive and negative emotions in a feeds as well as enthusiastic, intellectual, furthermore, basic parts of content dependent on the utilization of a lexicon containing words and their arranged classes. For instance, "concur" has a place with the word classifications: consent, full of feeling, positive feeling, positive feeling, and subjective process [9].

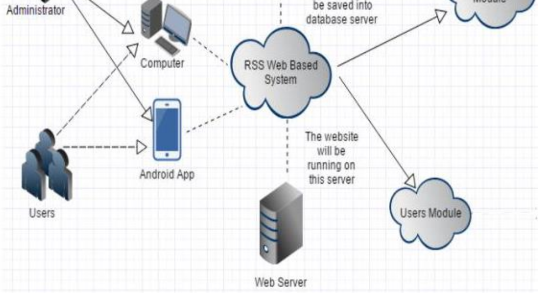
Database Server
Backup Database
Administrator Module will only be available to users who will have admin rights
Administrator
Administration Module
The news data will be saved into database server
Computer
RSS Web Based System
Android App
The website will be running on this server
Users
Users Module
Web Server

Fig.1. RSS Feeds E-Based System Architecture

## IV. Results and Discussion

Different RSS feeds data set has been analyzed to perform multiple sentiment analysis approaches using voyant-tools. The description of overall 500 RSS feeds (100 RSS feeds of each of the five sport categories) has been used for different experimental studies related to sentiment analysis. Initially, the RSS feeds collected from web or android applications through RSS feed news aggregators from various sources and categorized into five sports categories. The corpus contains the description of RSS feeds in the form of words and sentences for each sport category. The document comprises of feeds description is uploaded online on voyant-tools.org to perform different sentiment analysis experiments. The voyant-tools interfaces automatically generate using visualization tools such as word clouds, frequency graphs, summary, etc. upon completion of the upload.

Word clouds (also known as tag clouds or text clouds) have been generated with more specific and frequent words appear bigger and bolder in a source of the corpus. The purpose of word clouds can recognize patterns and examples that would some way or another are indistinct or hard to find in a tabular configuration. Frequently utilized keywords emerge better in a word cloud. Basic words that may be neglected in a tabular frame are highlighted in bigger content making them fly out when shown in a word cloud. Different word clouds of five sports groups RSS feeds using cirrus visualization tool in a voyant-tools have been produced are shown in Figs. 2, 3, 4, 5 and 6.

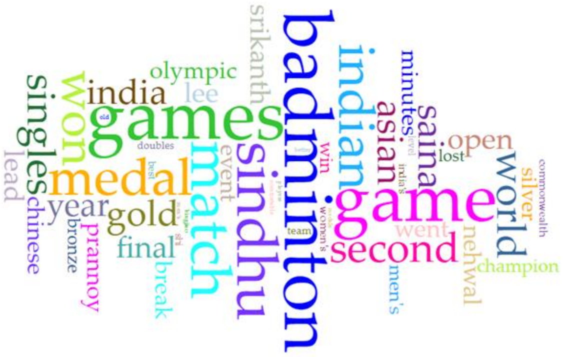

Fig.2. Batminton RSS Feeds Word Clouds

Fig. 2 clearly shows that the word badminton most frequently occurs in a corpus of batminton RSS feeds i.e., 24 times according to the automatic generated summary, then the word games occurs 22 times, game occurs 21 times, sindhu occurs 19 times, match occurs 18 times and so on in a word cloud of batminton RSS feeds corpus

Fig. 3 clearly illustrates that the word india most frequently occurs in a corpus of cricket RSS feeds i.e., 28 times according to the automatic generated summary, then the word pakistan occurs 25 times, team occurs 23 times, cricket occurs 22 times, cup occurs 21 times and so on in a word cloud of cricket RSS feeds corpus.

Fig. 4 clearly explains that the word league most frequently occurs in a corpus of football RSS feeds i.e., 19 times according to the automatic generated summary, then the word goal occurs 18 times, win occurs 14 times, award occurs 13 times, won occurs 13 times and so on in a word cloud of football RSS feeds corpus.

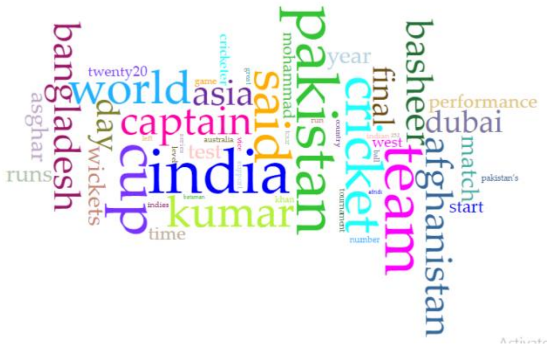

Fig.3. Cricket RSS Feeds Word Clouds

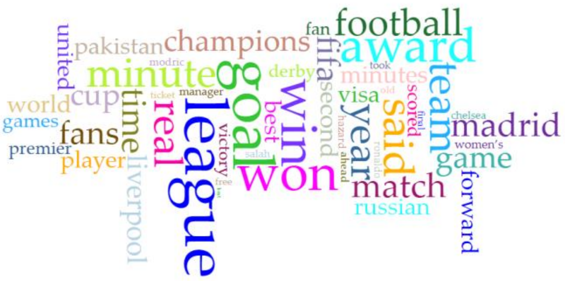

Fig.4. Football RSS Feeds Word Clouds

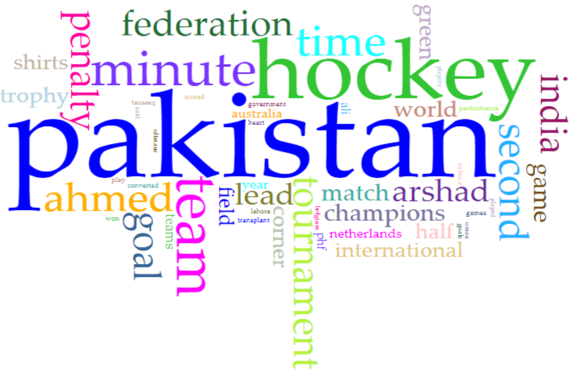

Fig.5. Hockey RSS Feeds Word Clouds

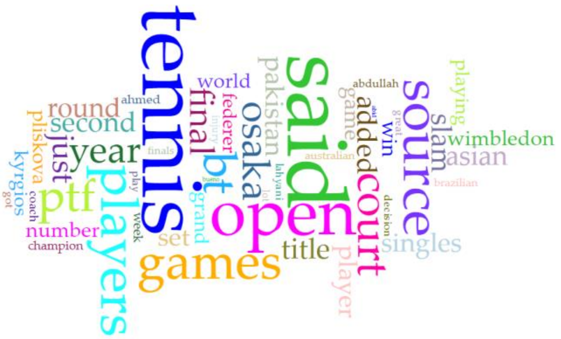

Fig.6. Tennis RSS Feeds Word Clouds

Fig. 5 clearly describes that the word pakistan most frequently occurs in a corpus of hockey RSS feeds i.e., 35 times according to the automatic generated summary, then the word hockey occurs 19 times, team occurs 12 times, minute occurs 11 times, time occurs 10 times and so on in a word cloud of hockey RSS feeds corpus.

Fig. 6 clearly defines that the word tennis most frequently occurs in a corpus of tennis RSS feeds i.e., 27 times according to the automatic generated summary, then the word said occurs 26 times, open occurs 17 times, players occurs 14 times, source occurs 14 times and so on in a word cloud of tennis RSS feeds corpus.

The stop wordlist option is also applied to generating word clouds. Stop words are basically an arrangement of ordinarily make use of words in a language. The reason behind why stop words are elementary to various applications is that, in the event where the words can be removed which are generally employed in a given language, can center on the essential words.

The term frequency graphs are generated using voyant-tools online for each sport category is shown in Figs. 7, 8, 9, 10 and 11. The purpose of term frequencies chart is the distribution of an occurrence of the word(s) in a form of frequencies in an entire corpus either using relative frequencies or raw frequencies for a word split into segments.

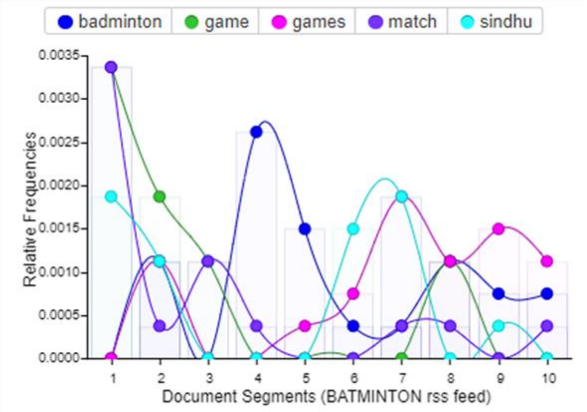

Fig.7. Batminton RSS Feeds Frequency Graph

Fig. 7 clearly shows that the word badminton has 25% relative frequency in a document segment 4, 15% relative frequency in a document segment 5, 5% relative frequency in a document segment 6, and 8% relative frequency in a document segment 9 and 10 of badminton RSS feeds and so on for other words whereas each document segment contains the text description of 10 RSS feeds.

Fig. 8 clearly illustrates that the word india has 7% relative frequency in a document segment 1, 20% relative frequency in a document segment 2, 18% relative frequency in a document segment 3, 7% relative frequency in a document segment 6, 10% relative frequency in a document segment 7, 20% relative frequency in a document segment 8, and 10% relative frequency in a document segment 9 of cricket RSS feeds and so on for other words.

Fig. 9 clearly elucidates that the word league has 13% relative frequency in a document segment 3, 23% relative frequency in a document segment 4, 5% relative frequency in a document segment 5 and 6, 17% relative frequency in a document segment 7, and 10% relative frequency in a document segment 9, of football RSS feeds and so on for other words.

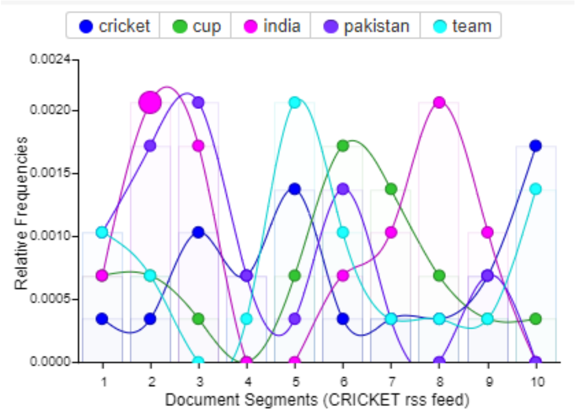

Fig.8. Cricket RSS Feeds Frequency Graph

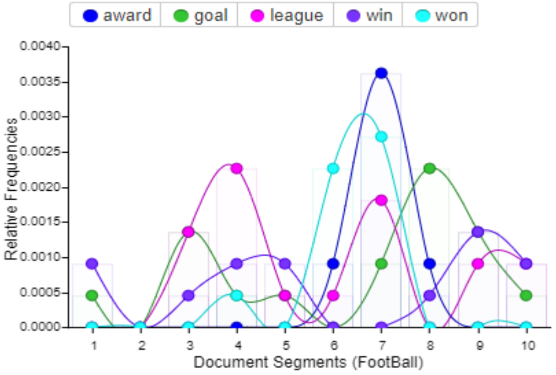

Fig.9. Football RSS Feeds Frequency Graph

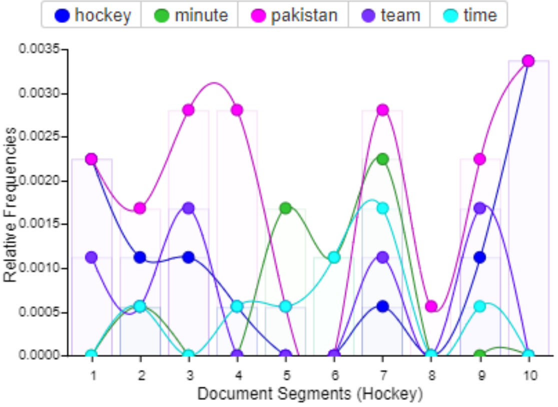

Fig.10. Hockey RSS Feeds Frequency Graph

Fig. 10 clearly interprets that the word pakistan has 22% relative frequency in a document segment 1, 18% relative frequency in a document segment 2, 20% relative frequency in a document segment 4 and 5, 20% relative frequency in a document segment 7, 5% relative frequency in a document segment 8, 18% relative frequency in a document segment 9 and 35% relative frequency in a document segment 10 of hockey RSS feeds and so on for other words.

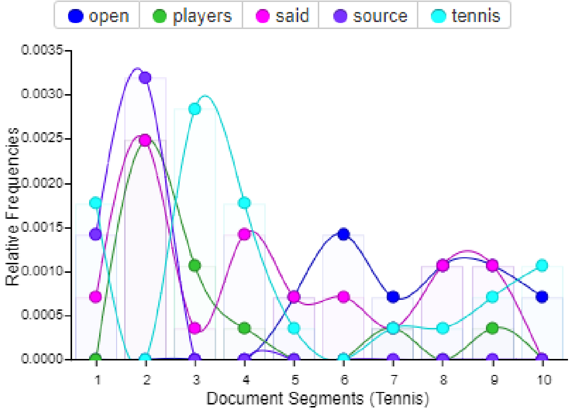

Fig.11. Tennis RSS Feeds Frequency Graph

Fig. 11 clearly interprets that the word tennis has 18% relative frequency in a document segment 1, 28% relative frequency in a document segment 3, 18% relative frequency in a document segment 4, 5% relative frequency in a document segment 5, 5% relative frequency in a document segment 7 and 8, 8% relative frequency in a document segment 9 and 10% relative frequency in a document segment 10 of tennis RSS feeds and so on for other words.

Scatter plots are produced for each of the sports categories are depicted in Figs. 12, 13, 14, 15 and 16 to show the correspondence of words used in a corpus. A cluster of words in the number of groups is automatically determined by the analysis criteria in a voyant-tool. Words in a group would show a proportion of similarity between the words. Clusters of terms will be seen as a single color.

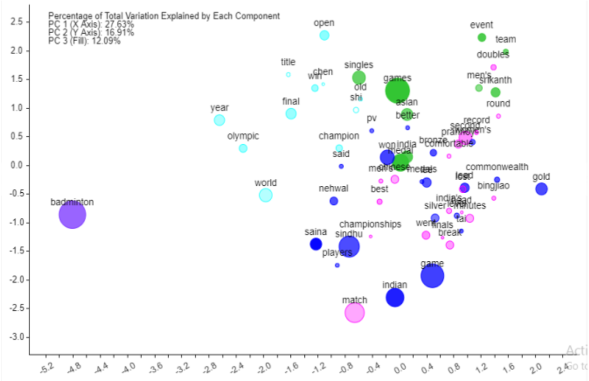

Fig.12. Batminton RSS Feeds Scatter Plot

Fig. 12 indicates that the word badminton forms the cluster with the highest raw frequency of 24 according to correspondence analysis and similarly the clusters of other words have formed accordingly with different raw frequencies of badminton RSS feeds.

Fig. 13 specifies that the word india forms the cluster with the highest raw frequency of 28 according to correspondence analysis and similarly the clusters of other words have formed accordingly with different raw frequencies of cricket RSS feeds.

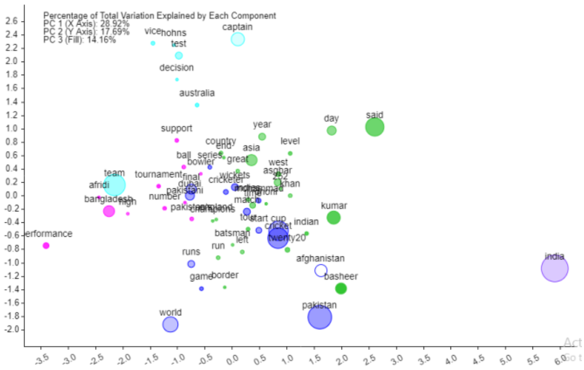

Fig.13. Cricket RSS Feeds Scatter Plot

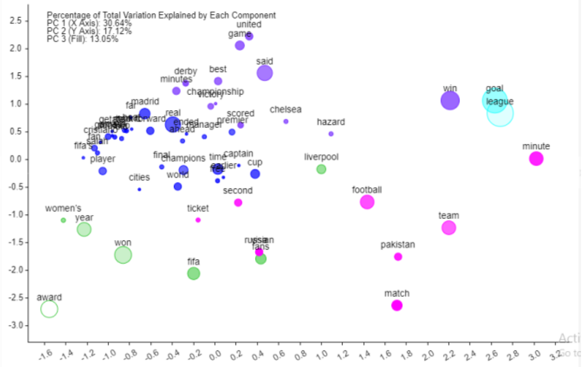

Fig.14. Football RSS Feeds Scatter Plot

Fig. 14 indicates that the word league forms the cluster with the highest raw frequency of 19 according to correspondence analysis and similarly the clusters of other words have formed accordingly with different raw frequencies of football RSS feeds.

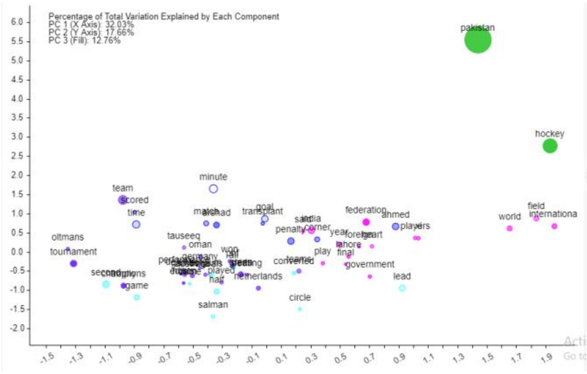

Fig.15. Hockey RSS Feeds Scatter Plot

Fig. 15 shows that the word pakistan forms the cluster with the highest raw frequency of 35 according to correspondence analysis and similarly the clusters of other words have formed accordingly with different raw frequencies of hockey RSS feeds.

Fig. 16 shows that the word tennis forms the cluster with the highest raw frequency of 27 according to correspondence analysis and similarly the clusters of other words have formed accordingly with different raw frequencies of tennis RSS feeds.

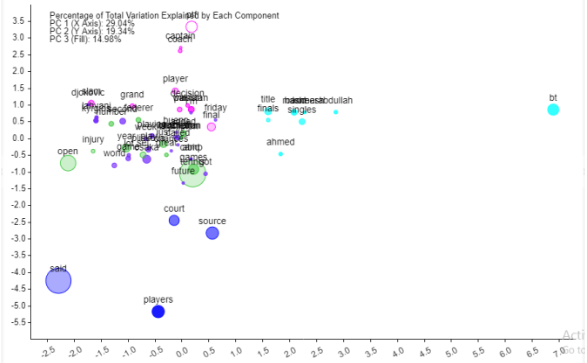

Fig.16. Tennis RSS Feeds Scatter Plot

According to voyant-tools analysis summary, the badminton RSS feeds corpus has 2,676 total words and 859 unique word forms with 32.1% vocabulary density and 22.1% average words per sentence. The cricket RSS feeds corpus has 2,915 total words and 1,071 unique word forms with 36.7% vocabulary density and 25.8% average words per sentence. The football RSS feeds corpus has 2,211 total words and 900 unique word forms with 40.7% vocabulary density and 27.0% average words per sentence. The hockey RSS feeds corpus has 1,784 total words and 682 unique word forms with 38.2% vocabulary density and 20.7% average words per sentence. The tennis RSS feeds corpus has 2,822 total words and 1,068 unique word forms with 37.8% vocabulary density and 23.1% average words per sentence.

Different emotions categorized as positive and negative have been automatically determined using LIWC online of sports RSS feeds description. Pie charts in Figs. 17, 18, 19, 20 and 21 are depicting the calculated sum of RSS feeds with a percentage of positive and negative emotions in feeds of each sports category.

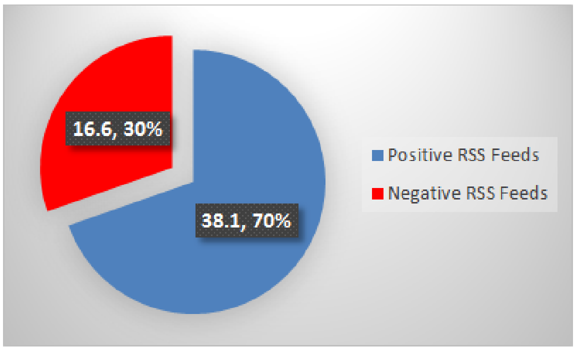

Fig.17. Batminton RSS Feeds Emotions

The column chart in Fig. 22 is depicting the overall summary of positive and negative emotions as a polarity associated with RSS feeds belonging to five different sports group produced by using LIWC tool. As shown in Fig. 22, Football has the highest positive polarity and Hockey has the least positive polarity. Similarly, Cricket has the highest negative polarity and Football has the least negative polarity. The purpose of polarity classification in a sentiment analysis under the perspective of this study is to determine the impact of the content in the description of RSS feeds for different sports on a reader/user-perceived as a positive emotion or a negative emotion respectively at the document level.

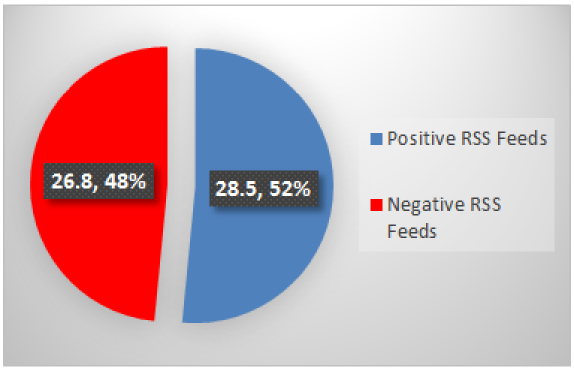

Fig.18. Cricket RSS Feeds Emotions

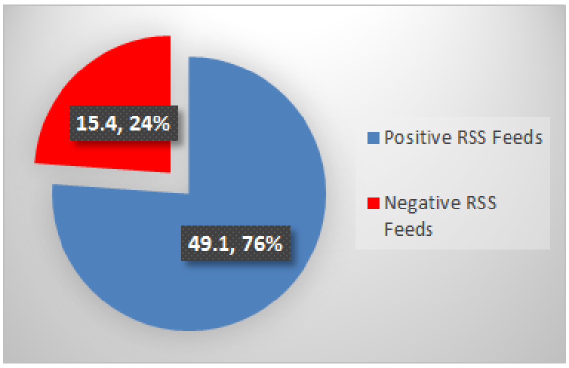

Fig.19. Football RSS Feeds Emotions

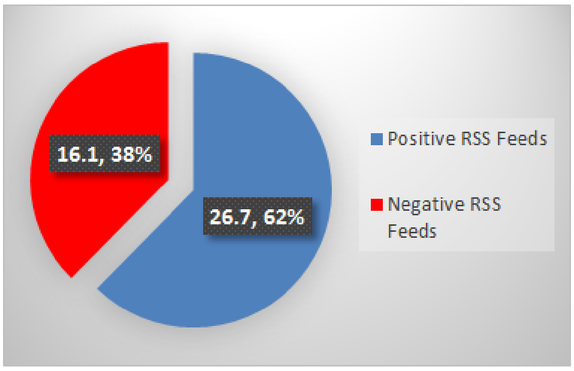

Fig.20. Hockey RSS Feeds Emotions

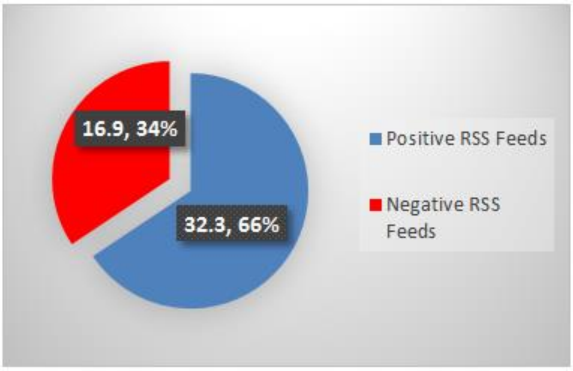

Fig.21. Tennis RSS Feeds Emotions

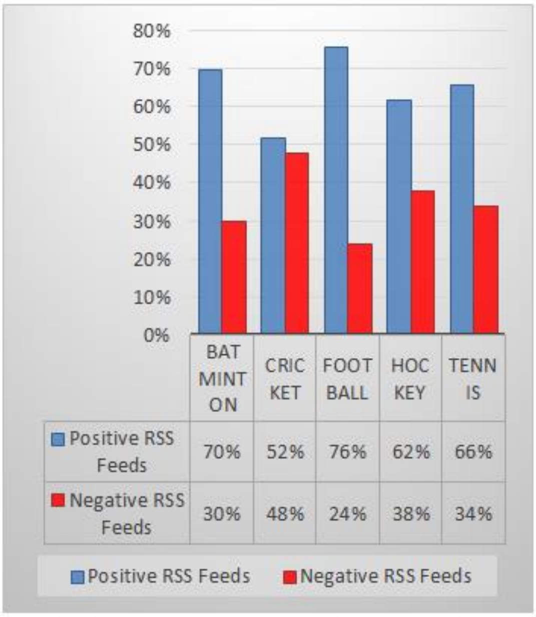

Fig.22. Overall Polarities in a Sport RSS Feeds

Additionally, the two statistical tests (F-test and Chi-Square) have been performed to test the following hypotheses:

Ho: Sports RSS feeds have an emotional impact on web users/readers
Ha: Sports RSS feeds does not have an emotional impact on web users/readers

F statistics are the ratio of two variances or two mean squares. The mean square is a normal dispersion, taken into account in the degree of freedom (DF), which is used to estimate the variance as shown in equation 1.

$$F=\frac{variation\ between\ RSS\ feeds\ values}{variation\ with\ in\ the\ RSS\ feeds\ values} \qquad (1)$$

The values obtained from LIWC for positive and negative emotions in RSS feeds are used to perform both F-test and Chi-Square method. The ANOVA summary and results of an f-ratio test for the treatments of RSS feeds are described below in Table 2.

Table 2. ANOVA Summary and Results obtained from f-ratio test

| **Summary of RSS Feeds Computation** | | | | |
|---|---|---|---|---|
| | ***Treatments*** | | | |
| | **Positive Emotions** | **Negative Emotions** | **Total** | |
| **N** | 5 | 5 | 10 | |
| **∑X** | 174.7 | 91.8 | 266.5 | |
| **Mean** | 34.94 | 18.36 | 26.65 | |
| **$\sum X^2$** | 6430.85 | 1775.78 | 8206.63 | |
| **Std.Dev.** | 9.0392 | 4.7522 | 11.0775 | |
| **Result Details** | | | | |
| ***Source*** | ***SS*** | ***df*** | ***MS*** | |
| **Between-treatments** | 687.241 | 1 | 687.241 | *F* = 13.1793 |
| **Within-treatments** | 417.164 | 8 | 52.1455 | |
| **Total** | 1104.405 | 9 | | |

Table 3. Chi-Square Computation

| **Chi-Square Method Results** | | | | | | |
|---|---|---|---|---|---|---|
| | Batminton | Cricket | Football | Hockey | Tennis | ***Row Totals*** |
| Positive Sentiments | 38 (36.05) [0.11] | 29 (36.70) [1.62] | 49 (41.95) [1.19] | 27 (28.18) [0.05] | 32 (32.12) [0.00] | 175 |
| Negative Sentiments | 17 (18.95) [0.20] | 27 (19.30) [3.08] | 15 (22.05) [2.26] | 16 (14.82) [0.09] | 17 (16.88) [0.00] | 92 |
| ***Column Totals*** | 55 | 56 | 64 | 43 | 49 | **267 (Grand Total)** |

The value of f-ratio test obtained is 13.1793. The p-value is .006682. The result is significant at *p* <.05, where the *p*-value is the probability to determine if F-value is frequent or rare is the assumption that the true null hypothesis. If the probability is low enough, it can be concluded that the data is not compatible with null hypotheses.

In the summary of Table 2. N=5 represents the five sports categories for both positive and negative emotions. The 500 different values for each RSS Feed obtained using LIWC are accumulated and the mean of the values is 34.92 and the standard deviation is found 9.0392 for positive emotions in RSS feeds. Likewise, the mean of the values is 18.36 and the standard deviation is found 4.7522 for negative emotions in RSS feeds.

The Table 3 presents the results obtained using Chi-Square method. When the Chi-Square method is applied to the same values give the statistics 8.5861. The *p*-value is.072319, where the *p*-value is the probability of obtaining such a chi-square is strong when two variables are independent. The result is not significant at *p* < .05. This test applies when two practical variables from a number are used to determine if there are significant relationships between two variables as shown in equation 2.

$$\chi^2 = \frac{(RSS\,feeds\ observed\ values\text{-}RSS\,feeds\ expected\ values)^2}{RSS\,feeds\ expected\ values} \tag{2}$$

For the f-ratio test, since .006682 < .05 the decision is not to reject the null hypothesis Ho. The data provided sufficient evidence at a 5% level of significance to conclude that the sports RSS feeds impact emotionally on web users/readers. On the other hand, for the Chi-Square test, since .072319 $\nless$ .05 the decision is to reject null hypothesis Ho. The data is not provided sufficient evidence at a 5% level of significance to conclude the dependency of sports RSS feeds impact emotionally on web users/readers.

## V. Conclusion

Text mining and sentiment analysis rise as a testing field with many impediments as it includes natural language processing. It has a wide assortment of utilizations that could profit by its outcomes, for example, news inspection, showcasing, question replying, readers or users do. Essential experiences from opinions are communicated on the web particularly from social media is indispensable for some organizations and institutions, regardless of whether it is as far as feedback on product, open state of mind, or financial specialists sentiments.

Sometimes information is estimated as unstructured and unorganized in a pre-characterized way. The greater part of this originates from content information, similar to messages, bolster tickets, chats, social media, studies and surveys, articles, and records. These texts are typically troublesome, tedious and costly to dissect, comprehend, and deal with. Sentiment analysis enables organizations to understand this unstructured information by automating data processing like processing RSS feeds.

In this study, a lexicon corpus-based approach is used on five different sports group RSS feeds taken from different news websites (obtained through web and android test application) and sentiment analysis is applied on the description of 500 RSS feeds containing text and phrases to perform various text analytics. Two online sentiment analysis and text mining tools are used to apply multiple sentiments analysis approach at document level of each sport category feed such as badminton, cricket, football, hockey, and tennis.

Different word clouds, frequency graphs, scatter plots, and summary have been generated using online open-source web-based voyant-tools to carry out RSS feeds analysis. Another online web-based tool used in this study is LIWC (Linguistic Inquiry and Word Count) to classify RSS feeds on each sport as positive and negative to identify the significant influence of news feeds about different sports on users/readers. It was found by analyzing the text in a description of RSS feeds that Football has the highest positive polarity i.e. 49.1 and Hockey has the least positive polarity i.e. 26.7. Similarly, Cricket has the highest negative polarity i.e. 26.8 and Football has the least negative polarity i.e. 16.1.

Moreover, the results of hypothesis from two different statistical tests were differing from each other in terms of *p*-value obtained after computation. The f-ratio test computation proven that the sports RSS feeds have an emotional impact on web users / readers at a 5% level of significance. On contrary, the chi-square test does not prove that the sports RSS feeds have an emotional impact on web users / readers at a 5% level of significance. It may depend on the situation as well as on the individual that either the things impact on reader’s mind or not or either they extract positivity or negativity from the content according to their interest.

However, more improved system can be built with dynamic data sets containing RSS feeds in different fields like education, business, medical, political, etc., where different documents features changes according to time and to prepare training data all over again with new features will be a crucial step in advance of classifying news feeds.

## VI. Future Work and Recommendations

The sentiment analysis has emerged as a challenging field, with numerous constraints, such as related to natural language processing. The challenge of this domain is to advance the machine's ability to understand the text that readers do. Reaching an important conclusion from the views expressed on the internet, especially from social media, is important, whether it is about product comments, public sentiments, or investor feedback.

The main idea of this research is to enhance the communication in advance by providing an RSS system which provides a web-based solution and which will automate the manual work. Even though, the sentiment analysis of the feeds on sports have been performed using different web tools to find out useful text patterns because the news deeply influence on readers or people in order to keep themselves update and use to with the information and particularly to identify the news hold either a positive impact or negative impact on readers mind.

However, the research needs to address some limitations in future related to the data size limit need to be enhanced to deal plentiful of data feeds. Another issue is the possible misclassification of RSS feeds text such as humor, irony or sarcasm terms or phrases based on a collection of positive and negative labeled feeds because words have different meanings and interpretations. So, it needs more additional details and strong techniques or algorithms of the aspects liked and disliked.

## VII. Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

## References

[1] J.A. Morente-Molinera, et al, "Analysing Discussions in Social Networks Using Group Decision Making Methods and Sentiment Analysis." *Information Sciences*, vol. 447, 2018, pp. 157–168, doi:10.1016/j.ins.2018.03.020.

[2] https://en.wikipedia.org/wiki/Sentiment_analysis [Aug. 29, 2018].

[3] https://en.wikipedia.org/wiki/RSS [Oct. 3, 2018].

[4] Y. Haribhakta, K.S. Doddi, "Categorization of News Articles using Sentiment Analysis." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. 2017 Sept&Oct; 2(5): 52–60.

[5] M. O'Shea, et al, "Mining and Visualising Information from RSS Feeds: a Case Study." *International Journal of Web Information Systems*, vol. 7, no. 2, 2011, pp. 105–129, doi:10.1108/17440081111141763.

[6] K.D. Gaikwad, et al, "Opinion Mining and Sentiment Analysis Techniques: A Recent Survey." *International Journal of Engineering Sciences & Research Technology* 2016 Dec; 5(12): 1003–6.

[7] http://www.rss-specifications.com/rss-faqs.htm [Oct. 4, 2018].

[8] S.V.S. Bharathi, A. Geetha, "Sentiment Analysis for Online Stock Market News using RSS Feeds." *International Journal of Current Engineering and Scientific Research*. 2017; 4(4): 58–63.

[9] A. D'Andrea, F. Ferri, P. Grifoni, T. Guzzo, 9. "Approaches, Tools and Applications for Sentiment Analysis Implementation." *International Journal of Computer Applications*. 2015 Sep; 125(3): 26–33.

[10] R. Kent Wills, "Efficient Sentiment Analysis of Feeds for Rapid User Information Gain." : 1–4.

[11] A.S. Dulange, R.B. Kulkarni, S.S. Ambarkar, "Opinion Mining From Blogosphere for Analysis of Social Networking." *International Journal of Soft Computing and Engineering*. 2013 Sep; 3(4): 141–146.

[12] J.M. Ruiz-Mart ńe, R. Valencia-Garc á, F. Garc á-Sánchez, "Semantic-Based Sentiment analysis in financial news.":38–51.

[13] S. Bharathi, A. Geetha, "Sentiment Analysis for Effective Stock Market Prediction." *International Journal of Intelligent Engineering and Systems*. 2017; 10(3): 146–54., doi: 10.22266/ijies2017.0630.16.

[14] E. Haddi, X. Liu, Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis." *Procedia Computer Science*. 2013; 26–32., doi:10.1016/j.procs.2013.05.005.

[15] https://www.w3schools.com/xml/xml_rss.asp [Oct. 5, 2018].

[16] https://en.wikipedia.org/wiki/News_aggregator [Oct. 5, 2018].

[17] https://voyant-tools.org/ [Oct. 7, 2018].

[18] https://tribune.com.pk/story/ [Sep. 30, 2018].

[19] https://www.samaa.tv/sports/2018 [Sep. 30, 2018].

[20] https://www.geo.tv/latest/ [cited 2018 Sep 30].

[21] https://www.dawn.com/news [cited 2018 Sep 30].

[22] https://www.thenews.com.pk/print/ [cited 2018 Sep 30].

[23] http://liwc.wpengine.com/ [cited 2018 Oct 10].

[24] https://www.newscientist.com/ [cited 2018 Oct 16].

[25] https://www.ukessays.com/essays/internet/ [cited 2018 Oct 16].

## Authors' Profiles

**Khalid Mahboob** earned master's degree in Computer Science & I.T. in 2011 from NED University of Engineering & Technology, Karachi, Pakistan. Currently enrolled in a Ph.D program at NED University of Engineering & Technology, Karachi, Pakistan. He is a Lecturer in Computer / Software Engineering Department at Sir Syed University of Engineering & Technology, Karachi, Pakistan. His research interests include, Educational Data Mining, Machine Learning, Expert System and Decision Support System, Sentiment Analysis.

**Fayyaz Ali** earned masters of science degree in Software Engineering in 2013 from University of Hertfordshire – UK. He is a Lecturer in Computer Science Department at Sir Syed University of Engineering & Technology, Karachi, Pakistan. His research interests include Software Engineering and Development, Software Methodologies, Artificial Intelligence, Project Management, Image Processing, and Sentiment Analysis.

**Hafsa Nizami** received her masters degree in Computer Science in 2016 from Sir Syed University of Engineering and Technology, Karachi, Pakistan. She is working as a Lecturer in Sir Syed University of Engineering and Technology, Karachi, Pakistan. Her research interests in Software Engineering, Cloud Computing, Open Source Software, Database Technology, Artificial Intelligence, Computer Science, and Big Data.

**How to cite this paper:** Khalid Mahboob, Fayyaz Ali, Hafsa Nizami, "Sentiment Analysis of RSS Feeds on Sports News – A Case Study", International Journal of Information Technology and Computer Science(IJITCS), Vol.11, No.12, pp.19-29, 2019. DOI: 10.5815/ijitcs.2019.12.02