Journal of Research in Business & Social Science 5 (1), 2016: 1-14

Research in Business and Social Science

IJRBS ISSN: 2147-4478

Contents available at *www.ssbfnet.com/ojs*

Doi:

Gossip Management at Universities using Big Data Warehouse Model Integrated with a Decision Support System

Pelin Vardarlıer

School of Business and Management Science, Istanbul Medipol University, Beykoz, İstanbul, 34810, Turkey.

Gökhan Silahtaroğlu

School of Business and Management Science, Istanbul Medipol University, Beykoz, İstanbul, 34810, Turkey.

Abstract

Big Data has recently been used for many purposes like medicine, marketing and sports. It has helped improve management decisions. However, for almost each case a unique data warehouse should be built to benefit from the merits of data mining and Big Data. Hence, each time we start from scratch to form and build a Big Data Warehouse. In this study, we propose a Big Data Warehouse and a model for universities to be used for information management, to be more specific gossip management. The overall model is a decision support system that may help university administraitons when they are making decisions and also provide them with information or gossips being circulated among students and staff. In the model, unsupervised machine learning algorithms have been employed. A prototype of the proposed system has also been presented in the study. User generated data has been collected from students in order to learn gossips and students' problems related to school, classes, staff and instructors. The findings and results of the pilot study suggest that social media messages among students may give important clues for the happenings at school and this information may be used for management purposes.The model may be developed and implemented by not only universities but also some other organisations.

Key Words: *Big Data, Data Collection, University Management, Gossip*

JEL classification: *M150, M1, C800, C550.*

Introduction

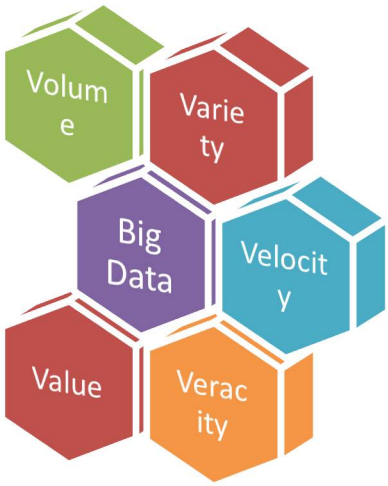
Modern business managements are benefitting from merits of digital era more and more. With the rapid improvement of computer devices, software tools and storage capacities, managers are much more comfortable to use these tools. Everyday another solution or idea emerges to assist modern managers when they are making their decisions. In this sense, Decision Support Systems (DSS) have been very helpful for years. Nevertheless, along with popularity of social networks and handy mobile devices, data stored in digital format have become enormous in size. Therefore, these days, valuable information hidden in this data has become much larger in volume than it ever was in the past. Data mining is to extract valuable information from large data sets, on which decision support systems are built. Thus, the importance of building data warehouses from Big Data and using it for management purposes is getting more and more crucial for competition, marketing, customer satisfaction and similar management issues. Managing a university is not only arranging classes, monitoring instructors and motivating them to work, teach and produce harder. Students are another important part of universities that should be taken into account for a better management. Their motivation and satisfaction is as important as that of instructors, research assistants and other office workers. That's why university administrators perform some surveys with students to learn how happy they are with campus facilities, teaching quality etc. However, this can be done only once or twice in an academic year. Although this is held to detect and solve or fix a problem, it is often too late to fix it, because the problem probably arose some time ago, and there is nothing to do for the person who suffered from it weeks, even months before it was detected. Yet, measures or precautions may be taken for future only. Therefore, it is obvious that university managers lack of a system that inform them about minor problems happening at school, misinformation among students, rumors and gossips. Using social media data integrated with other data available, a decision support system may be built to feed managers with information about minor problems that are happening somewhere in the university. Current technology, computer storage capacity, CPU speed, quality of algorithms and data available are altogether good enough to sort out this issue. In this study, a University Big Data Warehouse model which is enhanced with text mining and machine learning algorithms is presented.

In the first part of the paper we will give detailed information about Big Data, its key features, and challenges with Big Data and how to handle these challenges. In the second part, a literature review will be presented about how and to what extent Big Data and social media data are used for various management purposes. In the last part of the paper, you will find the proposed model and a sample study conducted within a prototype of the proposed system. Findings of the prototype study will also be discussed.

Literature Review

Big Data

Big Data is the collection of interrelated data gathered from different sources such as social media, transactional databases, maps, log files etc. Big Data is considered to be in the size of terabytes or petabytes. However, this is not the only feature of Big Data that makes it BIG. On the other hand, with this enormous size and being in different formats it has been a challenge to store, transfer and analyze Big Data. Big Data has five prominent characteristics or features. This is symbolized as 5 Vs of Big Data. As it is depicted in Fig 1. these characteristics are Variety, Velocity, Volume, Veracity and Value (Yu, 2013), (Zikopoulos, 2012).



Figire 1: Features of Big Data

Variety represents the data types that Big Data can be found in. As it is also stressed in the definition, Big Data may be generated in different formats and collected from various environments. For example, if it is map or geospatial data, it may be in the format of raster, vector or graph (Assuncao, Marcos D.et al, 2015) and if it is coming from social media it may be in the format of text and numbers residing in sentences or paragraphs. Variety feature of Big Data is considered to be high (Assuncao, Marcos D.et al, 2015). The second V for velocity refers to its being generated very fast and continuously. With the widespread and active usage of the Internet, users generate data every moment around the clock when it is thought globally. As companies and governments allow their customers and citizens to perform transactions online and when active social media usage is so popular, it goes without saying that data accumulation will speed up every day. This phenomenon will trigger the third V which stands for Volume. Perhaps it is the first thing to come to mind when Big Data term is first uttered. The adjective 'big' itself reminds us of high volume. Every day terabytes of interrelated data are generated and stored on disks. So, when we process Big Data we may be dealing with petabytes of data. Since Big Data is something in high volumes and coming from a variety of sources and formats, its quality is another important point. This is represented with Veracity. Although Big Data is thought to be authentic, it is very important that it ought to be valid and away from falsity. While millions of entries are stored in data disks each passing day, we cannot expect that all these are accurate and reliable data. Besides the provenance, integration of Big Data is another factor to determine the level of veracity. Even if the data are accurate and exact when they are put together or integrated for a Big Data Warehouse (BDW), their accuracy, reliability and exactness may be harmed because of disorganization. So, although all other Vs are thought to be high, when it comes to Veracity, it is doubted to be high and considered as low. The fifth V is Value. As it is seen, it will be a hard work to find, integrate, store and process this much data. It will need a great deal of effort. So, is that really worth it? Will you or your business be able to extract valuable enough information to satisfy the team and the cost sacrificed? It is not always possible to predict or forecast whether the data you will grab and tackle may yield valuable information that can be turned into benefits or not. Eventually, data selection and integration will play a crucial role for a future value to be gained. As the first V (Variety) suggests, data come from different sources. As it is depicted in Fig 2, some of the sources are as follows:

- Social media, such as Facebook, Twitter, LinkedIn, Pinterest, Google Plus+, Tumblr, Instagram, VK, Flicker etc include text, images and video files which are all stored and organized in free styles and formats.
- Corporate databases may be the most easily accessible data for businesses. They are stored and organized within Database Management Systems (DBMS) mostly with entity relational, object oriented or hierarchical data model (Ramakrishnan, R. and Gehrke J., 2003). All data are stored in properly labeled tables and connected to each other with certain ID numbers which are called primary or secondary keys.

- Another source for Big Data is databases managed and run by (mostly) state institutes. Criminal records, law system files and even health care system data are some of them. These data are stored and systematized or structured in DBMS like enterprise databases (Silberschatz. et al., 2010).
- Third party databases which are open to public usage are another important source. There are numerous web sites which store and display data such as, weather forecast, traffic information, finance, geospatial data (Lee, J. & Kang , M., 2015) and even DNA sequencing. Most of them may be considered as data with high veracity.
- Big Data across other organizations is merely another corporate database like Enterprise Data Warehouse (EDW) of a company or Big Data Warehouse which belongs to another institution.

Figure 2: Sources for Big Data

Although Big Data is considered to be as valuable as gold or petroleum, it is also as difficult as the others to be dealt with. Challenges to tackle Big Data may be categorized under four subtitles: Capturing, Storage, Integration and Processing.

Dealing with Big Data requires a big and well organized information system. As we all know, the very first thing an information system performs is to capture data (Hardcastle E., 2011). A regular Big Data Warehouse (BDW) will require a space of terabytes or petabytes, so it is not hard to imagine how difficult to capture it. That amount of data is mostly disk resident that cannot be transferred from one platform to another easily. In addition to this, the data do not have to be in only one, single place, platform or media. When the data reside in the transactional data base which belongs to a company, things are much easier. However, if it is in another company's or institution's DBMS, things will be more complicated in terms of accessing data. On the other hand, the data we are interested in may be at social media platforms or Big Data borrowed from other organizations. Then, difficulty level and type will change. This does not have to be because of technical issues. Problems may arise from ethical or confidential issues as well. After accessing data, we will face another question: how and where to store this huge amount of data. This is not only for space requirements. It is also for the structure of the system where data will reside. This is because, except DBMS , most of the data to compose BDW will be unstructured data (Chen, M. , 2014). Recently, some solutions have been developed to overcome this problem. Solutions are based on cloud technology as well as local databases.

Hadoop Distributed File System (HDFS) is a technology to overcome issues like capturing, storing and deployment (Chen, M., 2014). Hadoop is a framework that uses clusters of computers. It is an open source MapReduce implementation. HDFS partitions data sets and carries them on different nodes. MapReduce is

a programming model to process large amount of data which reside on clusters of computers (Chen et al., 2012), (Guo & Fox, 2012). The task Hadoop performs is to partition and replicate data chunks across multiple nodes and to provide APIs to be used with MapReduce applications (Guo & Fox, 2012).

HBase is another solution that can be used on top of Hadoop and HDFS (Bhupathiraju & Ravuri, 2014). It is and open source non-relational database model.

PIG Latin is a scripting language which enables Hadoop users to write MapReduce applications. It runs in YARN (MapReduce) and access any dataset stored in HDFS (Huang et al, 2012). As Hadoop and HDFS can be used in local drivers or dedicated servers there are plenty of cloud services which support HDFS and MapReduce processes. Google File System, Amazon Simple Storage Service, Nirvanix Cloud Storage, Open Stack Swift and Windows Azure Binary Large Object are some of them (Assuncao, Marcos D.et al, 2015).

There are also data warehouse software platforms developed for managing and querying large data sets which reside in Hadoop. One them is HIVE (Song et al, 2015). HIVE has a SQL like language called QL or HIVE QL. Although it is built under Hadoop project now it is a standalone project itself. SQL MapReduce is another platform to enable users to write MapReduce functions in programming languages such as Java, C#, Python, C++, and R (Gu et al, 2014), (Yuan et al, 2010), (Plimpton & Devine, 2011).

Empirical Review

Extracting information is the main focus in Big Data process. In literature there are plenty of tools and algorithms to analyze datasets and generate informative reports. Data analysis may be held in three different ways. It may be descriptive, predictive and prescriptive (Linda et al, 2015), (Zhang et al, 2001). Descriptive analysis answers the question of “What happened?”. For a descriptive analysis or information extraction, besides statistics, data mining algorithms are also used. K-means and derivatives of K-Means such as Spherical K-Means (Agarwal & Singh, 2012), Bisecting K-Means (Kashef & Kamel, 2009), (Savaresi & Boley, 2001), K-Means ++,(Agarwal & Singh,2012) also Suffix Tree Clustering (Han et al, 2006), (Largeron-Leténo, 2003) and Fuzzy Clustering [Son, 2015] algorithms are used for structured and text data. These algorithms are centroid based algorithms. On the other hand algorithms like DBSCAN, DENCLUDE,OPTICS and DBCURE-MR operate on the principle of density to segment data (Younghoon et al, 2014). All of them are widely used algorithms for a descriptive analysis. Another way of data analysis is Predictive Big Data analysis which deals with future. It tells what will happen and when it will happen. For predictive Big Data analysis, decision tree algorithms like C4.5 (Quinlan, J., 1993) , J48 (Bhargava, 2013), CART (Breiman et al., 1984), or Artificial Neural Network algorithms (Fausett A., 1994) are mostly used. All algorithms and techniques make predictions using the data available. The third type of Big Data analysis is prescriptive. Although Prescriptive Big Data analysis tells about future like predictive analysis do, it mostly answers questions like “why will it happen?” and it also gives answers to questions like how to prevent it happening. Furthermore this analysis may suggest possible and feasible options to exploit some future events. Algorithms from different disciplines like machine learning, computer vision, signal processing are used for prescriptive analysis (Bose et al., 2001) , (Vallmuur, 2015).

Integration is the fourth challenge to tackle. Since data are collected from a variety of sources and in different formats, it demands a big and reasonable endeavor to integrate or put them together. Integrating data of two or more different businesses stored in different systems is relatively easier, because in DBMS data are stored in a logical order. There are unique primary and foreign keys to identify each record, entity or relation. Data redundancy is not allowed in transactional databases. However, if it is a geospatial data or social network data containing images and text, there will be no predetermined attributes, tables or IDs to help you put them together as in DBMS. Matching two or more data entries or records in different sources is one challenge as organizing and integrating them is another. Often, there is no data model and meta data for the Internet data which makes most of Big Data.

Big Data has recently been used by many scholars from universities and also by researchers and technicians in different sectors. Tourism is one of them. Using web data, Fuchs et.al. presented a knowledge infra structure to be used for mountain tourism (Fuchs, 2014). They used business intelligence

approach within the framework of Destination Management Information System (DEMIS) to understand and assess customer behavior and experience before and after travelling. They used web navigation, booking and feedback data within DEMIS. Their proposal to use Big Data and business intelligence tools together will enable managers to receive real time information on tourists' on-site behavior and destinations. The Big Data which is made up of electronic word-of-mouth (Luo and Zhong, 2015) in social networks have been used by some other researchers to understand tourist behaviours and promote or manage tourism activities efficiently in Istanbul (Altunel and Erkut, 2015), Honk Kong(Vu, 2015), Finland (Mynttinen, 2015) and so on.

For marketing, there are plenty of studies. Using clickstream data, it is possible to extract online customers' age, gender and any other behavioral patterns (Silahtaroglu, 2015a). It is sort of web farming that researchers can harvest valuable information for managers to be used for marketing (Silahtaroglu and Dönertasli, 2015b). Xu et.al. studied the effects of using Big Data analysis to promote new products in the market. The study which has been held from a knowledge fusion perspective suggests that Big Data analytics play a prominent role on new product success (Xu, 2015). Another study shows that Big Data has the potential to impact nearly every area of marketing, and it is a new competitive advantage tool to be used by firms, therefore companies which will not build and employ a Big Data warehouse will face challenges in competition in the market (Erevelles, 2015).

Big Data residing on the Internet and social networks have been used at times of government crises for communication. The research have examined 300 local government officials from municipalities across the United States and samples of the crises like wildfires in California, the 2009 crash of U.S. Airways flight 1549, the 2010 Haiti earthquake, and Hurricane Sandy have been used, Graham (2015). Twitter and Facebook data have been used to analyse the role of user generated data during the school crisis on September 30, 2014 during active shooter incidents in a high school, in the USA, Mazer et al. (2015). When there is a crisis, thoughts, gossips, rumours, information and even misinformation started to be texted among people. The study presented a model that schools may manage social media during and after such crises to help control the situation. A similar model has also been proposed by Yates and Paquette after the Haitian Earthquake in 2010. They have showed how social media supports disaster and emergency response mechanisms from an organizational level, Yates and Paquette (2011). Two different reseach and study shows that a proper data warehouse of Big Data may be used for tobacco usage Link et al. (2015) and anti-smoking campaigns Chung (2015).

Nguyen et al., discussing the effect of web content, social networks and Big Data on brand innovation, stresses that customer needs can be retrieved as information via data mining and business intelligence tools, Nguyen et al. (2015). Researchers also emphasise that if social media content is transformed into a data warehouse and analysed, it will provide valuable information for competition and may be used for strategic behaviour by business managements. As it is also stressed in other studies Luo et al. (2015), Kim et al. (2015), Big Data analysis is an asset for firm value.

Big Data analysis has been used to discover whether there is a relationship between social media usage and academic performance of students as well. A research held in Saudi Arabia, Alwagait (2014) and another one in the US and Europe, Ozer (2014) are two of the recent studies. In addition a more specific study has been condected by Zhang et al. Exploiting social media content with the help of CiteSpace II, they showed that the evolution of social networks develop Collaborative Learning (CL) and eventually, Big Data accumulated in social networks may be used by instructors and university managements if it is handled properly Zhang (2015).

Research and Methodology

Proposed Model and Analysis

Big Data mostly come from social networks and web content. Young people and university students are probably the ones who use social networks most frequently. University students add a lot of data to social networks, so it is clear that they play an important role in developing Big Data. Since we know that there is very valuable and useful information hidden in Big Data why not using it for university management for

various purposes such as class allocations, facility organizations, rumor and gossip control, problem detection? In this study, we propose a data warehouse a model which enables to sort out some management issues via machine learning algorithms.

The model requires a dedicated server to keep the University BDW and perform the data retrieving, cleaning and analysis processes seamlessly. The university Big Data Warehouse is to consist of seven data sources. We propose them as follows:

- Social Network Data: This may be taken from one or more social networks such as, Facebook, Twitter, Pinterest, Instagram etc.
- University Student Database: This should include university automation system data like grades, attendance and so on.
- Learning Management System (LMS): announcements by instructors, tasks, assignments, lecture notes and comments may be included.
- University official mail server data: this may or may not include staff emails.
- Third party Data: Traffic information and weather information are some of them.
- Web content: Depending on the objective of the management one or more news portals may be added to be analyzed along with student and enterprise database.
- Office Automation System (OAS).

We have realized a prototype of the system with 300 students who are active Twitter users. We have built a database of those 300 students with the following data: Gender, year, GPA, school or department. A summary of this dataset is given in Table 1. 165 of volunteers were girls. Average GPA of students is 2.41. 50 students are freshmen and 40 of them are girls. 63 are sophomores and 31 of sophomores are girls, 32 are boys. 125 of all students were junior students and 72 were girls. Finally, we had 62 senior students; 22 girls and 40 boys.

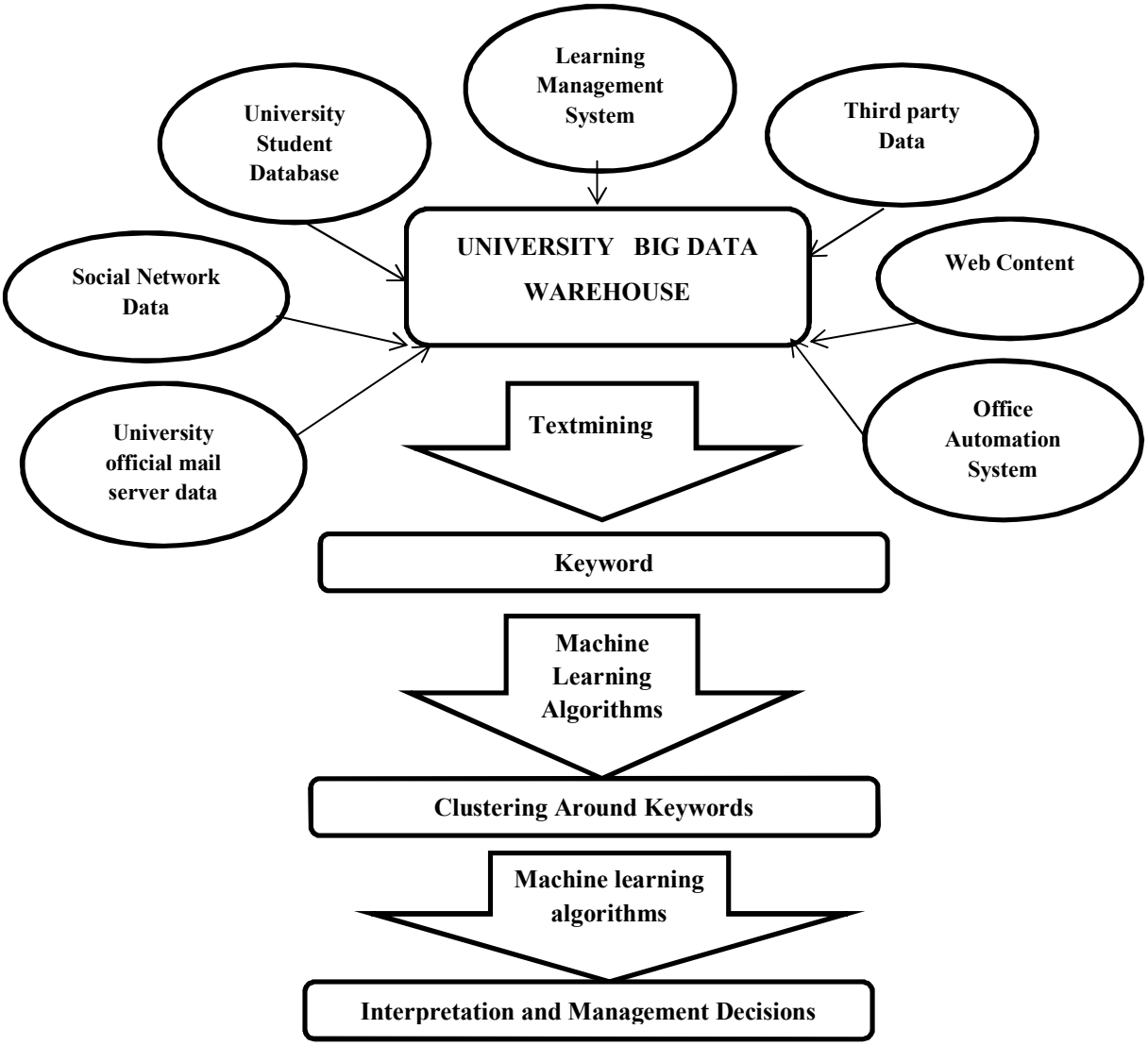


Figure 3: Overall Model proposed.

Empirical Data and Analysis

We have also used Application Program Interface (API) tools to grab the tweets of those volunteers. Moreover, they forwarded their messages related to school to an email address so that they could be stored in a separate database. All volunteers were motivated to use their Twitter and mail accounts actively to talk about subjects related to university.

Table 1: Summary of Student İnformation Dataset.

	Average/ Distribution	Female	Male
GENDER		165	135
YEAR	1(50);2(63);3(125);4(62)	1(40);2(31);3(72);4(22)	1(10);2(32);3(53);4(40)
GPA	2.41 (Average)	2.54	2.27

So, we had three datasets; one in database format, one in Excel format, Twitter text data and one email data in text format. We filtered the Twitter data with some key words and also with most frequently used words in a certain period of time. The keywords we used are the name of the university and the word "university" in a number of spelling formats. When we used the name of the university to collect keywords appearing, we were not limited with the tweets of 300 students. In addition, we have collected keywords within various time periods as much as Twitter allowed. This dataset has been united with excel data which

holds students information like gender, GPA etc. Secondly, we did the same filtering with email data and united the three data sets as shown in Fig 4 below.

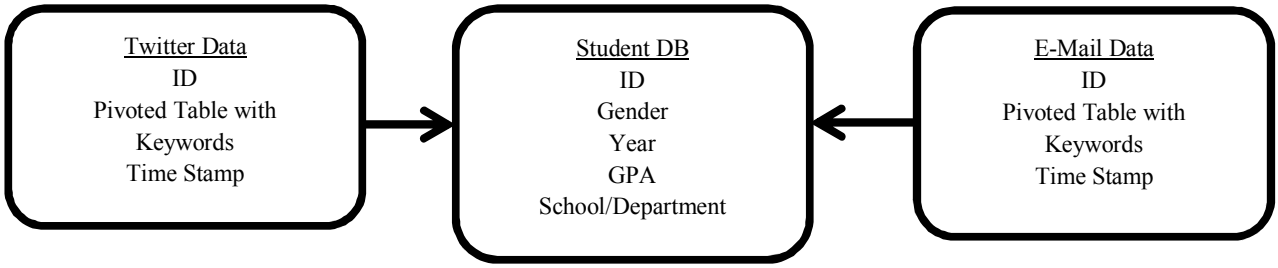


Figure 4: Pilot Big Data Warehouse

We used Artificial Neural Networks and DBSCAN algorithms to analyze and cluster the data around most frequently used words, keywords and timestamp. The algorithm created clusters around a keyword in the middle.

Results and Discussion

The analysis indicates that if some important words or phrases have been chosen by the model correctly, they may give valuable information about the happenings at school. The phrases automation system, the name of an instructor, holiday, lunch and elevator has been chosen as topics by the system. Each of these words or phrases makes up a cluster. The topic words are the cluster centers. Fig 5 represents one of the clusters. Results are informative enough to be used by university management. Here in Table 2, we present the most striking and usable keywords and summary of results in various time periods.

Table 2 Findings of the study.

Duration	Keyword	School	Gender	GPA	Year	Total (Users)
3 days	Automation System	No Majority	No Majority	2.2	1(30);2(20);3(15);4(2)	67
1 day	Instructor Mrs X.	Architecture	Majority Female	1.81	1(0);2(0);3(45);4(5)	50
5 days	Midterms	No Majority	No Majority	1.99	1(30);2(38);3(79);4(24)	171
3 days	Half Day (holiday)	Engineering, Business	No Majority	2.1	1(43);2(28);3(16);4(2)	89
10 days	Elevator					
25 Days	Lunch	GENERAL	KEYWORD			

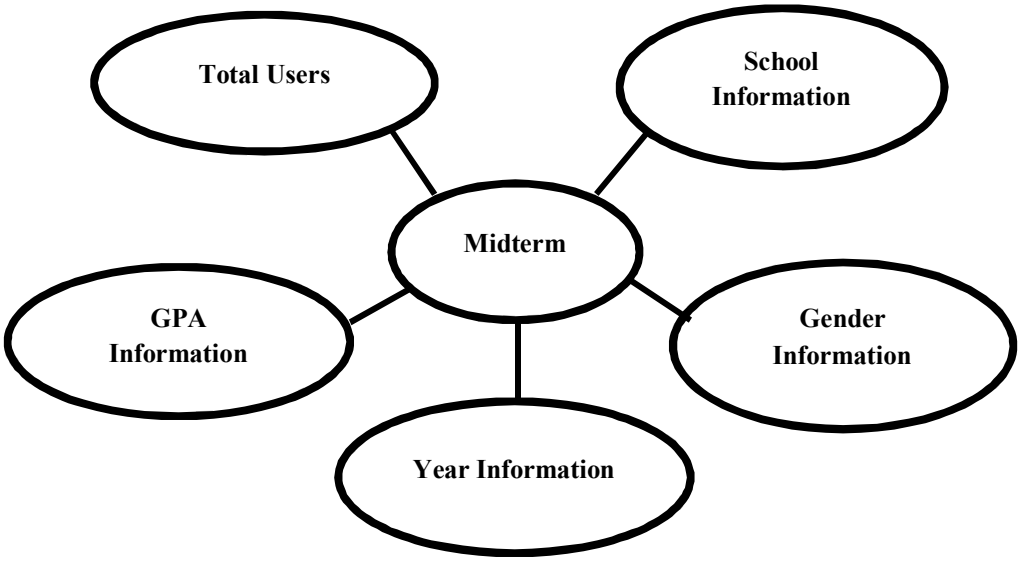


Figure 5: A prototype of clusters.

The pilot analysis shows that student automation system was mentioned by 67 students for 3 days. In fact, there was a problem with the student automation system during those dates. So, if the system had been used by the university management at that time, it could have been learned by the management easily and they would have had a chance to intervene with it. Here, we see that our decision support model enhanced with Big Data worked well. In the second row of Table 2, we see that 50 students talked about an instructor, some Mrs. X. Actually we did not check if there was a problem or not, yet it is obvious that there was an issue with that instructor. It may have been a problem or a piece of homework or assignment or any other gossip. During the midterm week, 171 students mentioned midterm and that is quite normal. Just before a Turkish national holiday, 89 students probably wanted to learn if there would be any classes on that day which is a half day off. We see that only engineering and business students were talking about this. We may guess that they were not informed about half day holiday classes. Last two rows have been taken from general keyword clustering which is done around the name of the university. We see that for 10 days *elevator* and 25 days *lunch* were hot topic among students. One can easily guess that there may have been a problem with the elevators and lunch at school.

Conclusion

Big Data has been used by practitioners and scholars and academics. In this study, we suggest a novel University Big Data Warehouse (UBDW) for universities to be used by university managements. The model consists of multiple data sources. These are social media data source which may be grabbed from social networks such as Twitter, Facebook, Pinterest, Instagram etc.; students' database source which includes gender, class, department, GPA etc.; students' school mail data source, learning management system (LMS) data source which includes lecture notes, comments on lecture notes, announcements by instructors etc.; students' automation system data which includes enrollment and registration data, grades, students mails to instructors and groups also surveys about classes and staff; office automation system data used by staff, third party databases and finally other web content.

Apart from this we built a prototype of the proposed system and tested it with 300 volunteer students. Students agreed to use their twitter and mail accounts actively for 65 days. They forwarded their mails which are related to school to an email address. We created a dataset with students' gender, year, GPA and school or department data. All these data were converted into a prototype Big Data warehouse after being transformed, filtered and cleaned. We applied some clustering process around most frequently used

words and university's own name. When we used the university' name as keyword we were not limited to the tweets of those 300 students, but we have collected and finally clustered all tweets between two certain dates. For clustering we used DBSCAN and Artificial Neural Networks algorithms. Our model generated twenty five clusters with a keyword in the center. However, we filtered and used only six of them which may be used by university management. Our model and study showed that during these 65 days of study, phrases *students' automation system, Instructor some Mrs. X*, *midterms*, *half day classes*, *elevator* and *lunch* have been discussed and gossiped by students. Our study also gives information about the students who talked about these subjects. The information supplied by the system or model are as follows: Majority of the gender of the students, average GPA of students who are talking on a certain issue, school or department of the students and how many years they have been attending the university.

When we examined and thought about the cluster center keywords we see that our model is informative enough to feed university management with information about issues that needs to be dealt with on the spot. Although our model may be used for different management purposes and tactical and strategical decision making processes, our prototype worked well for problem, gossip and rumor management. We believe that with the development of technology, faster, more informative and smarter systems will be built for the use of managers both at universities and other sectors.

Reference

Agarwal, S., Yadav, S., & Singh, K. (2012, March). Notice of Violation of IEEE Publication Principles K-means versus k-means++ clustering technique. InEngineering and Systems (SCES), 2012 Students Conference on (pp. 1-6). IEEE.

Altunel, M. C., & Erkut, B. (2015). Cultural tourism in Istanbul: The mediation effect of tourist experience and satisfaction on the relationship between involvement and recommendation intention. Journal of Destination Marketing & Management, 4(4), 213-221.

Alwagait, E., Shahzad, B., & Alim, S. (2014). Impact of social media usage on students academic performance in Saudi Arabia. Computers in Human Behavior.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 79, 3-15.

Bhupathiraju, V., & Ravuri, R. P. (2014, March). The dawn of Big Data-Hbase. In IT in Business, Industry and Government (CSIBIG), 2014 Conference on(pp. 1-4). IEEE.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. Information & management, 39(3), 211-225.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software. Pacific California.

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big Data: related technologies, challenges and future prospects. Springer.

Chen, Y., Alspaugh, S., Borthakur, D., & Katz, R. (2012, April). Energy efficiency for large-scale mapreduce workloads with significant interactive analysis. In Proceedings of the 7th ACM european conference on Computer Systems (pp. 43-56). ACM.

Chung, N., Lee, S., & Han, H. (2015). Understanding communication types on travel information sharing in social media: A transactive memory systems perspective. Telematics and Informatics, 32(4), 564-575.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904.

Fausett, Laurene. Fundamentals of neural networks: architectures, algorithms, and applications. Prentice-Hall, Inc., 1994.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big Data analytics for knowledge generation in tourism destinations–A case from Sweden. Journal of Destination Marketing & Management, 3(4), 198-209.

Graham, M. W., Avery, E. J., & Park, S. (2015). The role of social media in local government crisis communications. Public Relations Review.

Gu, R., Yang, X., Yan, J., Sun, Y., Wang, B., Yuan, C., & Huang, Y. (2014). SHadoop: Improving MapReduce performance by optimizing job execution mechanism in Hadoop clusters. Journal of Parallel and Distributed Computing,74(3), 2166-2179.

Guo, Z., & Fox, G. (2012, May). Improving mapreduce performance in heterogeneous network environments and resource utilization. In Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (Ccgrid 2012) (pp. 714-716). IEEE Computer Society.

Hardcastle Elizabeth, (2011). Business Information Systems, Elizabeth Hardcastle & Ventus Publishing ApS.

Huang, J., Ouyang, X., Jose, J., Wasi-ur-Rahman, M., Wang, H., Luo, M., ... & Panda, D. K. (2012, May). High-performance design of hbase with rdma over infiniband. In Parallel & Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International (pp. 774-785). IEEE.

il Han, S., Lee, S. G., Kim, K. H., Choi, C. J., Kim, Y. H., & Hwang, K. S. (2006). CLAGen: A tool for clustering and annotating gene sequences using a suffix tree algorithm. BioSystems, 84(3), 175-182.

Kashef, R., & Kamel, M. S. (2009). Enhanced bisecting k-means clustering using intermediate cooperation. Pattern Recognition, 42(11), 2557-2569.

Kim, S., Koh, Y., Cha, J., & Lee, S. (2015). Effects of social media on firm value for US restaurant companies. International Journal of Hospitality Management, 49, 40-46.

Kim, Y., Shim, K., Kim, M. S., & Lee, J. S. (2014). DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce. Information Systems, 42, 15-35.

Largeron-Leténo, C. (2003). Prediction suffix trees for supervised classification of sequences. Pattern Recognition Letters, 24(16), 3153-3164.

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. Big Data Research, 2(2), 74-81.

Link, A. R., Cawkwell, P. B., Shelley, D. R., & Sherman, S. E. (2015). An Exploration of Online Behaviors and Social Media Use Among Hookah and Electronic-Cigarette Users. Addictive Behaviors Reports.

Luo, N., Zhang, M., & Liu, W. (2015). The effects of value co-creation practices on building harmonious brand community and achieving brand loyalty on social media in China. Computers in Human Behavior, 48, 492-499.

Luo, Q., & Zhong, D. (2015). Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites. Tourism Management, 46, 274-282.

Mazer J.P et al. 2015." A cross-cultural qualitative examination of social-networking sites and academic performance", Computers in Human Behavior, 53, 238-248.

Miner, L., Bolding, P., Hilbe, J., Goldstein, M., Hill, T., Nisbet, R., ... & Miner, G. (2015). The Nature of Insight from Data and Implications for Automated Decisioning: Predictive and Prescriptive Models, Decisions, and Actions,Practical Predictive Analytics and Decisioning Systems for Medicine, Chapter 22, Pages 1008-1018.

Mynttinen, S., Logrén, J., Särkkä-Tirkkonen, M., & Rautiainen, T. (2015). Perceptions of food and its locality among Russian tourists in the South Savo region of Finland. Tourism Management, 48, 455-466.

Neeraj, D. B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), 3(6).

Nguyen, B., Yu, X., Melewar, T. C., & Chen, J. (2015). Brand innovation and social media: Knowledge acquisition from social media, market orientation, and the moderating role of social media strategic capability. Industrial Marketing Management.

Ozer, I., Karpinski, A. C., & Kirschner, P. A. (2014). A cross-cultural qualitative examination of social-networking sites and academic performance. Procedia-Social and Behavioral Sciences, 112, 873-881.

Plimpton, S. J., & Devine, K. D. (2011). MapReduce in MPI for large-scale graph algorithms. Parallel Computing, 37(9), 610-632.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. Morgan Kauffmann.

Ramakrishnan, R., & Gehrke, J. (2003). Database management systems, 3rd Edition,McGrwaw-Hill.

Savaresi, S. M., & Boley, D. L. (2001, April). On the performance of bisecting K-means and PDDP. In SDM (pp. 1-14).

Silahtaroglu G. (2015a) "Predicting Gender of Online Customer Using Artificial Neural Networks", 2nd International Conference on Management and Information Technology, pp. 45 – 50, New York, USA.

Silahtaroglu G., Donertasli H., (2015b), "Analysis and Prediction of E-Customers' Behavior by Mining Clickstream Data", Proceedings of 2015 IEEE International Conference on Big Data (Big Data), pp. 1466 – 1472, San Jose, USA.

Silberschatz, A. et al.(2010). Database System Concepts, McGrwaw-Hill.

Son, L. H. (2015). A novel kernel fuzzy clustering algorithm for geo-demographic analysis. Information Sciences: an International Journal, 317(C), 202-223.

Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., & Pierson, J. M. (2015). HaoLap: a Hadoop based OLAP system for Big Data. Journal of Systems and Software, 102, 167-181.

Vallmuur, K. (2015). Machine learning approaches to analysing textual injury surveillance data: A systematic review. Accident Analysis & Prevention, 79, 41-49.

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. Tourism Management,46, 222-232.

Xu, Z., Frankwick, G. L., & Ramirez, E. (2015). Effects of Big Data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. Journal of Business Research.

Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake.International Journal of Information Management, 31(1), 6-13.

Yuan, Y., Wu, Y., Feng, X., Li, J., Yang, G., & Zheng, W. (2010). VDB-MR: MapReduce-based distributed data integration using virtual database. Future Generation Computer Systems, 26(8), 1418-1425.

Zhang, L., Price, R., Aweeka, F., Bellibas, S. E., & Sheiner, L. B. (2001). Making the most of sparse clinical data by using a predictive-model-based analysis, illustrated with a stavudine pharmacokinetic study. European journal of pharmaceutical sciences, 12(4), 377-385.

Zhang, X., Wang, W., de Pablos, P. O., Tang, J., & Yan, X. (2015). Mapping development of social media research through different disciplines: Collaborative learning in management and computer science. Computers in Human Behavior.

Zikopoulos, P., & Eaton, C. (2012). Understanding Big Data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.