Data Mining as a tool for Knowledge Discovery

Fotini Kolokathi, Vasilis Stavrou

{p3090088@dias.aueb.gr , stavrouv@aueb.gr}

Information Security and Critical Infrastructure Protection (INFOSEC) Laboratory

Dept. of Informatics, Athens University of Economics & Business (AUEB), Greece

Introduction

- Rapid explosion of Online Social networks has turned users from passive receivers into integral parts of them.
- Users transfer their offline behavior to the online world.
- Extraction of new information from users data contributes to the profiling of them.

Subject of research

- Application of data mining techniques to messages of users who intend to commit suicide.
- Purpose: Knowledge Discovery which is useful to other sciences such as Psychology.

Data Preprocessing

- Creation of a web service that:
  - converts greeklish into Greek.
  - remove noise from data such as intonations in greek words, punctuations etc.

	id	year	month	day	time	anartisi	anartisi_cleaned	meso	location	age	fulo	entopismos	katastasi
Edit Copy Delete	121	2011	10	14	09.00	x8es eftasa sto terma... to plisiasa... simera den tha milagame... enw... amaeixa figei...ola tha hxan teliosei...	χθες εφτασα στο τερμα το πλησιασα σημερα δεν θα μιλαγαμε ενω amaeixa φυγειολα θα ειχαν τελειωσει	FACEBOOK		0	?	?	?
Edit Copy Delete	122	2011	10	14	11.00			SMS	ΠΕΡΙΣΤΕΡΙ/ ΑΤΤΙΚΗ	0	m	yes	abnormal
Edit Copy Delete	123	2011	10	16	09.00	...8A ΑΥΤΟΧΤΟΝΙSW. ΕΧ8ΕS ΤΟ MESIMERI IPIA MIA GOYLIA FITOFARMAKO...	ΘΑ ΑΥΤΟΧΤΟΝΙΣΩ ΕΧΘΕΣ ΤΟ ΜΕΣΗΜΕΡΙ ΗΠΙΑ ΜΙΑ ΓΟΥΛΙΑ ΦΥΤΟΦΑΡΜΑΚΟ	FACEBOOK		0	?	?	?
Edit Copy Delete	124	2011	10	16	11.00	Αποφάσισα σήμερα να αυτοκτονήσω έχω βαρεθεί να με προσβάλει της κάθε πουτάνας ο γιος θα ήθελα λοιπόν ως τελευταία χάρη να ζητήσω την οριστική διαγραφή μου από το site το συντομότερο δυνατόν	Αποφασισα σημερα να αυτοκτονησω εχω βαρεθει να με προσβαλει της καθε πουτανας ο γιος θα ηθελα λοιπον ως τελευταια χαρη να ζητησω την οριστικη διαγραφη μου απο το site το συντομοτερο δυνατον	www.myphone.gr	ΝΕΑ ΧΩΡΑ ΧΑΝΙΩΝ	35	m	yes	normal

Figure_1:Database fields

Statistical diagrams

- Diagrams arised from users data that are stored in database.

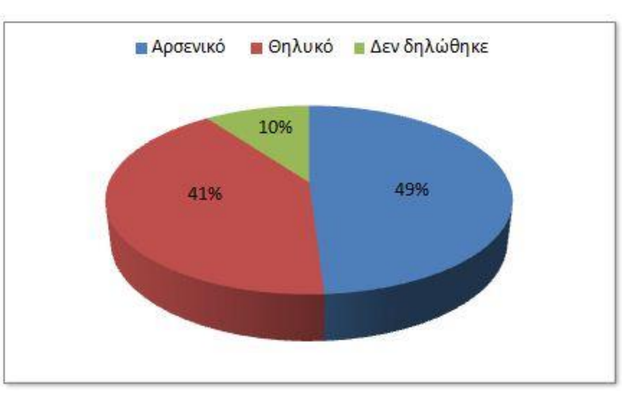

Figure_2: Gender

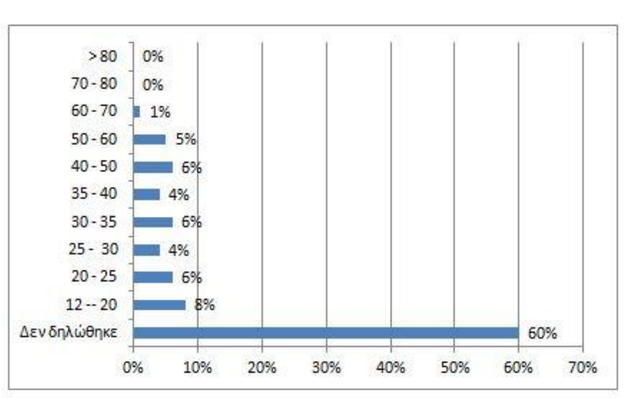

Figure_3: Age

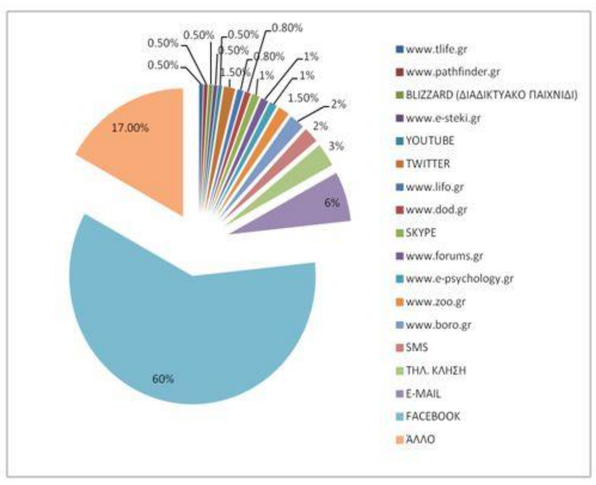

Figure_4: Sources of users messages

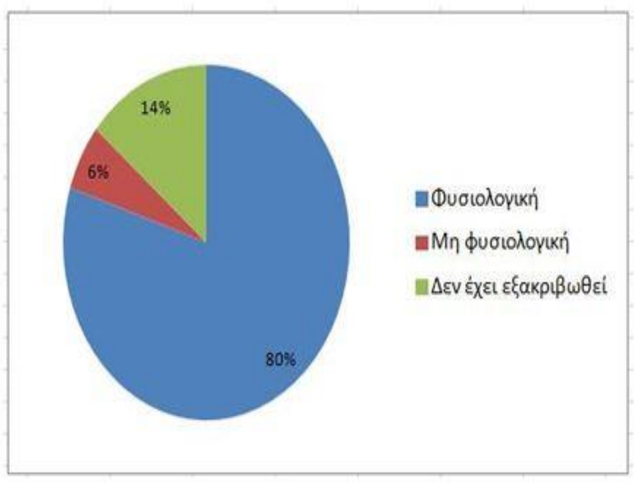

Figure_5: State of users

- Comparison between users who were found in normal and abnormal state.

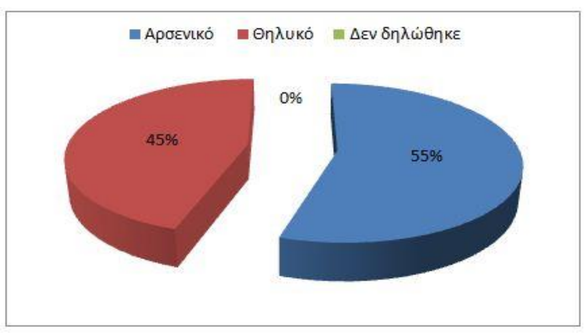

Figure_6: Gender of users who were found in normal state

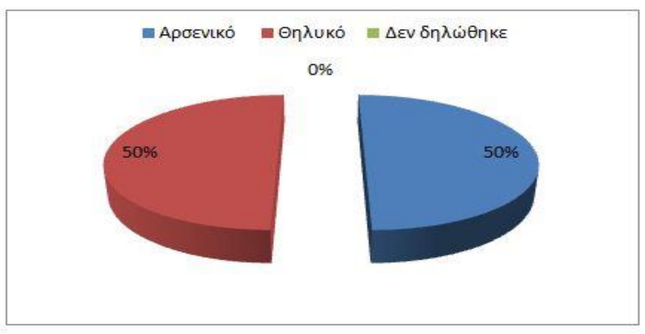

Figure_7: Gender of users who were found in abnormal state

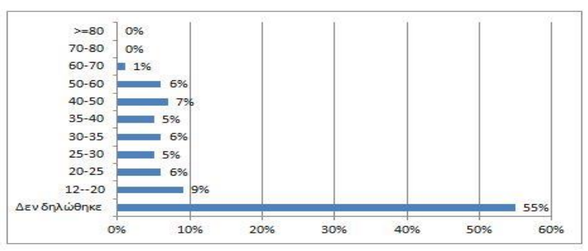

Figure_8: Age of users in normal state

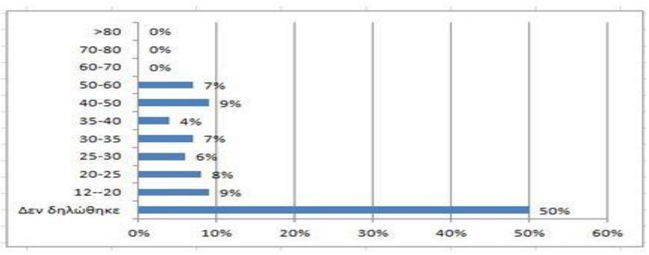

Figure_9: Age of users in abnormal state

Data clustering

- Clustering in Weka:
  - Comparison between clusterers.
  - Clustering of texts/phrases via SimpleKMeans.

Clusterer	Number of iterations	Sum of squared errors	Time to build model(sec)	Incorrectly clustered instances
MakeDensityClusterer	4	6211	0.07	36.9369%
SimpleKMeans	4	6211	0.03	36.8243%
FilteredClusterer	4	6211	0.05	36.8243%

Table_1: Performance of clusterers

- Tag cloud generation for each cluster.

Figure_12: Tag cloud for cluster_1

Figure_13: Tag cloud for cluster_2

Usage Patterns

- In the context of this research, usage patterns are phrases with great frequency in the collected texts.
- Extraction of usage patterns from texts.

Figure_14: Tag cloud for patterns

Association rules

- Mining association rules between sets of items in database.
  - Extraction of obvious results due to the multiple empty fields of the database.

```
1. anartisi_cleaned=          627 ==> anartisi= 627    <conf:(1)> lift:(1.42) lev:(0.21) [164] conv:(184.29)
2. anartisi= 627 ==> anartisi_cleaned=          627    <conf:(1)> lift:(1.42) lev:(0.21) [164] conv:(184.29)
3. anartisi_cleaned=          entopismos=yes 563 ==> anartisi= 563    <conf:(1)> lift:(1.42) lev:(0.19) [165] conv:(165.48)
4. anartisi= entopismos=yes 563 ==> anartisi_cleaned=          563    <conf:(1)> lift:(1.42) lev:(0.19) [165] conv:(165.48)
5. anartisi_cleaned=          katastasi=normal 513 ==> anartisi= 513    <conf:(1)> lift:(1.42) lev:(0.17) [150] conv:(150.78)
6. anartisi= katastasi=normal 513 ==> anartisi_cleaned=          513    <conf:(1)> lift:(1.42) lev:(0.17) [150] conv:(150.78)
7. anartisi= katastasi=normal 513 ==> entopismos=yes 513    <conf:(1)> lift:(1.14) lev:(0.07) [62] conv:(62.39)
8. anartisi_cleaned=          katastasi=normal 513 ==> entopismos=yes 513    <conf:(1)> lift:(1.14) lev:(0.07) [62] conv:(62.39)
9. anartisi_cleaned=          entopismos=yes katastasi=normal 513 ==> anartisi= 513    <conf:(1)> lift:(1.42) lev:(0.17) [150] conv:(150.78)
10. anartisi= entopismos=yes katastasi=normal 513 ==> anartisi_cleaned=          513    <conf:(1)> lift:(1.42) lev:(0.17) [150] conv:(150.78)
```

Conclusions

- Most users, who expressed intention to commit suicide in their messages,:
  - had not revealed their ages and were males in their majority.
  - expressed this via Facebook.
  - were using particular words and phrases such as death, end, suicide, " end of my life" etc.
- Users who were found in normal state had similar characteristics with users who were found in abnormal state.
- Extraction of reliable conclusions for users who were found in abnormal state is not feasible due to the low rate of their database records.

References

1. Amichai-Hamburger, Y., Vinitzky, G.," Social network use and personality. In: Computers in Human Behavior", vol. 26, pp. 1289-1295, 2010.
2. Bishop M., "Pattern Recognition and Machine Learning", Springer, 2007.
3. Boyd, D., "Social network sites: Public, private, or what. Knowledge Tree", vol. 13, no. 1, pp. 1-7, 2007.
4. De Choudhury, M., Counts, S., "The nature of emotional expression in social media: Measurement, inference and utility", Human Computer Interaction Consortium (HCIC) Workshop, 2012.
5. Gritzalis D., Kandias M., Stavrou V., Mitrou L., "History of Information: The case of Privacy and Security in Social Media", in Proc. of the History of Information Conference, Law Library Publications, Athens, 2014.
6. Kandias M., Stavrou V., Bozovic N., Mitrou L., Gritzalis D., "Can we trust this user? Predicting insider's attitude via YouTube usage profiling", in Proc. of 10th IEEE International Conference on Autonomic and Trusted Computing, pp. 347-354, IEEE Press, Italy, 2013.
7. Kandias M., Mitrou L., Stavrou V., Gritzalis D., "Which side are you on? A new Panopticon vs. privacy", in Proc. of the 10th International Conference on Security and Cryptography (SECRYPT-2013), pp. 98-110, Iceland, 2013.
8. Kandias M., Galbogini K., Mitrou L., Gritzalis D., "Insiders trapped in the mirror reveal themselves in social media", in Proc. of the 7th International Confe-rence on Network and System Security (NSS-2013), pp. 220-235, Springer (LNCS 7873), Spain, June 2013.
9. Kandias M., Stavrou V., Bozovic N., Mitrou L., Gritzalis D., "Can we trust this user? Predicting insider's attitude via YouTube usage profiling", in Proc. of 10th IEEE International Conference on Autonomic and Trusted Computing (ATC-2013), pp. 347-354, IEEE Press, Italy, 2013.
10. Liu B., "Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data", Springer, 2007.
11. Mitrou L., Kandias M., Stavrou V., Gritzalis D., "Social media profiling: A Panopticon or Omniopticon tool?", in Proc. of the 6th Conference of the Surveillan-ce Studies Network, Spain, 2014.
12. Oreilly, T., "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software", Communications & Strategies, No. 1, p. 17, 2007.
13. Pallavi , Godara S., "A Comparative Performance Analysis of Clustering Algorithms", Vol. 1, Issue 3, pp.441-445, 2011.
14. Singhal S., Jena M., "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", 2013.
15. Siriporn O., Benjawan S., "Anomaly Detection and Characterization to Classify Traffic Anomalies", 2008.
16. Witten I., Frank E., "Data Mining Practical Machine Learning Tools and Techniques", Elsevier, 2005.