I.J. Intelligent Systems and Applications, **2015, 04, 17-25**
Published Online March 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.04.03

Automatic Ethical Filtering using Semantic Vectors Creating Normative Tag Cloud from Big Data

Ahsan N. Khan
Teradata Corporation, Lahore Office, Pakistan
E-mail: ahsan.nabi@gmail.com

***Abstract*—**Ethical filtering has been a painful and controversial issue seen by different angles worldwide. Stalwarts for freedom find newer methods to circumvent banned URLs while generative power of the Internet outpaces velocity of censorship. Hence, keeping online content safe from anti-religious and sexually provocative content is a growing issue in conservative countries in Asia and The Middle East. Solutions for online ethical filters are linearly upper bound given computation and big data growth scales. In this scenario, Semantic Vectors are applied as automatic ethical filters to calculate accuracy and efficiency metrics. The results show a normative tag cloud generated with superior performance to industry solutions.

***Index Terms*—**Semantic Vectors, Censorship, Distributional Semantics, Normative Systems, Tag Cloud.

I. Introduction

Ethical filtering has traditionally been manual blacklisting of IPs at the DNS, ISP and HTTP level in conservative countries in Asia and The Middle East like Pakistan (Nabi, 2013), Azerbaijan (Nazirova, 2012) and Syria (Chaabane et al, 2014). Although they don't cover all sites in question, other methods of filtering have problems of their own. Filtering requests based on keyword analysis traps false positives and has caused collateral damage in Syria, denying proxies for legitimate use. Still, social networks could not be censored except for a few specific pages.

Instant Messaging like Skype and Hangout have different censorship levels, and so do proxies have varying tolerance for Tor networks. Complete site blockage of youtube.com and proxy sites in Pakistan is spurned with discontent as primitive and encroaching internet freedom. Still, Virtual Private Networks, torrents, web/socks proxies and Tor onion networks are examples of circumvention consistently evolved to evade new censorship standards. Social Networks, despite encouraging violations of norms, have survived censorship through group user moderation. Facebook, for instance, has inadvertently given indulgences for consumer 'stare', sometimes marginalizing the norms of privacy and decency (Veer, 2011), but still moderated and managed social acceptability.

Ethical filtering focuses mainly by the collective concerns of human rights, IT and Web compliance, and social norms. Browsers and search engines bind parental controls to prevent child pornography, cyber-crimes, drug-trafficking and contraband items. Software programs work by scanning web-site addresses, web-site content, email and other documents to find objectionable words or concepts. Filter levels of these software programs range from most restrictive to least restrictive. In case a valid website is blocked by software error, often a painstaking process of application is undertaken to open up the website, consulting principal or higher authority, writing application, filling out forms, and requesting access rights through software exception handling mechanisms.

Ethical decision making is a specialized form of decision making, where businesses and management concerns and training programs can separate their range of choices based on ethical constraints. Ethical Usage is also being taught in academic environments. As is said with concern "With Power Comes Responsibility", the missing link in the ethical filtering is we need to define global and nation-level issues like privacy, security, international freedom and justice (Diebert et al, 2008).

From 2008, OpenNet Initiative started publishing surveys of global internet filtering regimes. China was among the first countries to adopt national filtering system, a firewall on the main Internet chokepoints, ISPs and gateways, 'The Great Firewall of China'. This first generation of Internet censorship denied access based on IP lists, keywords and domains. Main targets of censorship included child pornography, terrorism and cyber security.

With the second generation of control, normative systems of overt and covert tracks legalized controls by specifying conditions and terms-of-use, and "just in time" control in volatile situations, e.g. during election campaigns. Third generation of control no longer denied access, but countered threats by campaigns of repairing information. This involved expansive surveillance, data mining, viral attacks to campaign sites, distributed denial-of-service attacks, legal notices to take down sites, and national information-sharing strategies (Diebert et al, 2010).

II. Problem Statement

The wider domain of censorship we encounter in the Internet, the larger the chance of false positives hurdling free access, learning and growth in cyberspace. Another problem is the amount of textual content becoming

available every year matches the level of computing power, so Moore's Law may not be relevant for big data analytics without faster or at least linear order text processing algorithms (Widdows and Cohen, 2010). To cater for this, an automated information-sharing strategy needs to be developed. Focus needs to be shifted from eliminating blacklists to understanding whitelists for efficient rating on growing graylists automatically. In this study, we do automatic document classification learning from blacklists and whitelists to rate and classify graylists on a linear scale of efficiency.

In the next section, we see some related work in solving the problem of ethical filtering using network architecture, categories, keywords and document classification. In Section IV we define our methodology for classifying and clustering documents based on ethical feature vectors. In Section V we list down and analyze the results of using our methodology. Finally we conclude on the preference of using Semantic Vectors for automatic filtering and generating normative tags online on the cloud computing scale.

III. Related Work

A. Cyber Code of Ethics

Lawrence Lessig in his several books on the influential effects of media, culture, and economy on Cyber Code, has pointed to some questions like 'Should the hate speech, political speech and such online be regulated by the state?', 'How to strike a balance between user privacy and law-enforcement of intellectual property, business and trade secrets, ideas and expression?', 'Is the way it is the way it must be?', 'What things regulate?', 'If the Internet can't be regulated, why?' (Lessig, 2006) Lessig and his teammate Jonathan Zittrain points out that the original TCP/IP architecture of the Internet allowed for its anonymity, generative networks within networks allowed for innovation and disruption, and 'appliancized' proprietary networks allowed for heightened regulability (Zittrain, 2008). As solutions in Web 2.0, Zittrain emphasize role of Internet Service Providers (ISPs) to use community based filtering tools, and the PC manufacturers and protocol designers to bridge the gap of information divide that caused anonymity in the Internet. Legal governance is questioned on all end-points, but social solutions like Wikipedia are given as an alternative. These online solutions for collaborative information verification, spam blocking and reporting offensive content involve community participation and work better for social problems than laws.

B. Filtering levels over networks

Censorship in Pakistan is chiefly manual blocking of IPs, IP ranges, specific files, file types or folders, domains and subnets serving offensive and objectionable content related to blasphemy, anti-religious, and adult-rated content. Proxies and circumvention sites were also blocked. Up to 50 million URLs can be blocked in 1ms latency. Blacklists can be maintained in databases. However, blacklists are maintained manually, with no use of keyword filtering within URL or content to learn new URLs. The problem here is growing the lists automatically keeping in pace the generative power of the Internet. (Nabi, 2013)

Censorship in Syria is done in increasing levels of control from keywords, strings, categories, IPs, subnets, and domains. There are categories (custom or default) assigned based on URL request. Custom categories had moderated redirection policies, while the majority of filtering was done by default categories. Default included DENIED, where filtering is fully done or PROXIED, where some nodes could be allowed access as an exception.

Custom categories had specific policies of redirection from blocked contents in Online Social Networks like Facebook and Twitter. Targeted censorship was performed based on keywords in content (e.g. proxy, Israel), or keywords in URL cs-host, cs-path or cs-query fields. Manually specified lists (e.g. 'Syrian Revolution' page on Facebook) were also blocked. Some domains were filtered like Netlog and Badoo while entire subnets like 84.229.0.0/16, and domains of countries like .il were blocked in Syria. (Chaabane et al, 2014)

In a survey of the most popular filtering mechanisms (Leberknight et al, 2012), operational costs, accuracy and granularity increase linearly from using (1) IP filters (2) DNS filters (3) keyword filters to (4) stateful traffic analysis including Deep Packet Inspection (DPI). Most filtering software is developed internally, while some content filtering solutions like SmartFilter and deep packet analyzers are commercially provided by Secure Computing, Nokia, Siemens, etc. Standard internet filtering software like Barracuda, CyberPatrol, FilterGate, and Websense provide accuracy of 87% for direct URL access of adult content and 81% accuracy for keyword searches of adult content. Filtered searches of content not adult-oriented have correspondingly lesser accuracy, a low 67% for keyword searches. Hence filtering terrorism, anti-state or hate material by keyword searches has considerably low accuracy (Jan, 2008).

C. Countering Anti-censorship technology

Anti-censorship tools like Tor and Hotspot can be controlled. Web proxies and virtual private networks (VPNs) create SSL-based encrypted HTTP tunnel and redirect traffic through their relays. Application-level tunnels can be detected by statistical fingerprinting (Dusi at al, 2009). Filtering HTTP requests from Tor offer simple technical problem of regular expression match, while Tor-Onion requests offer the challenge of decryption. Still some inconsistent circumvention of encrypted Tor can be managed by identifying Tor bridges and hence cracking the IPs associated with the onion network (McLachlan and Hopper, 2009). As in Pakistan, the general layman community is already not familiar with use of public VPNs and Web proxies, by the time technical skills for circumvention becomes available, the censorship policy can mature (Nabi, 2013).

D.Filtering by topics and keywords

China's micro blog site Weibo has an automated censorship mechanism based on keywords searching, backwards repost searching, specific users monitoring, search query filtering, timeline filtering and user rating systems. Weibo keeps more than one keyword filter lists and deletes posts that contain any of the keywords, while closely monitoring the topics and the authors of the deleted posts. It also keeps tracks of the search queries containing the author, topic or keywords of the deleted post. This way, the censorship mechanism automatically filters the prolific post generation. There are 70,000 new posts generated per minute would be too expensive to be filtered manually, hence automated (Luhn, 1958).

Automatic topic extraction, proposed originally in 1958 (Salton and Buckley, 1988), used words weighted by their frequency and other statistical measurements on phrasal, sentence and paragraph level. In micro blogs and instant messaging, traditional natural language processing could not be employed because grammatical structures, spellings, and native vocabulary are not used. Two text mining approaches have rather been much popular: N-grams and TF*IDF (Song et al, 2012). For cases where words and meanings do not map one on one, like in Chinese micro blogs, the trigrams of characters are taken as a point in reconstructing aggregate meanings. Such modern technique of information retrieval is referred to as Pointillism [22] and is particularly effective in detecting deviation from regular lexicon typically found in the social media.

Distributional semantics, the empirical study of meanings of words found in large-text corpora, takes the term frequencies and inverse document frequencies (TF*IDF) to find out term and document similarities that can be used in classifying and retrieving terms and documents in question. The underlying model is geometric representing probability distributions of words as vectors in high dimensional spaces. Starting with the term-document matrix, the sparseness is dealt with by singular value decomposition and random projection to come up with reduced pseudo-orthogonal term vectors and document vectors. Cosine similarities for these vectors are calculated more accurately and faster, hence scaling up linearly with increasing online corpora. These are utilized in Semantic Vectors package (Widdows and Cohen, 2010), which we would be using for building ethical filters in our methodology.

IV. Methodology

First we collect blacklists of URL and keywords. Then we devise strategy to learn more keywords of similar patterns. Third we classify and create clusters of internet content into whitelists, graylists and blacklists.

The whitelists will contain none of the keywords in the blacklists and minimal of the keywords matched in the graylists. Graylists may contain one or two blacklist keywords, but document must show farthest conjunction with blacklists, meaning distance between graylists and blacklists is to be maximized by removing documents that maximize blacklist keywords and including the complement: keywords common to the whitelist.

For this we installed Semantic Vectors Package 5.4 on Eclipse Luna with Maven m2e integration plugin, and Java Development Kit 1.8. We linked to Maven repository, included dependencies of lucene-demo-4.9.0, lucene-analyzers-common-4.9.0, lucene-expressions-4.9.0, lucene-facet-4.9.0, lucene-queries-4.9.0, lucene-queryparser-4.9.0, jsoup and junit.

Then we built our lucene index of search terms by reading corpus using the commands:

```
java org.apache.lucene.demo.IndexFiles
-index ${workspace_loc}\src\indexfiles
-docs
"${workspace_loc}\src\test\resources\testdata
\[CORPUS_PATH]"
java pitt.search.semanticvectors.BuildIndex
-luceneindexpath
${workspace_loc}/src/indexfiles    -vectortype
real positional_index/
-filternumbers false
-docindexing incremental
```

For the corpus, we used seven sets, comprising mainly online blacklist documents of anti-religious and sexual nature from developing and developed countries, graylist documents containing religious debates and reddit comments, whitelist documents containing scriptures and academic documents, and unrated documents of newsgroups taken as control experiment. The two categories (1) anti-religious and (2) sexual had to be taken care of separately for training indexes and search results.

We found blacklist documents from the following lists for Pakistan, Denmark and Australia. We used banned word list keywords from four sources to search for sexually explicit material in Australian/Danish blacklisted content for verification of our indexes, and searched more in unrated content for relative comparison. If our search in blacklisted content showed significantly higher match rate than the control group of unrated content, we would then test the graylist of social reddit comments for similarity measures related to whitelist and blacklist and determine which list is closer to the reddit comments.

Similar trends we would find for religious/anti-religious content. We would learn indexes from Pakistani blacklists. However, the problem in finding anti-religious comments is that there is no fixed vocabulary set typifying anti-religious content. Hence we would use Semantic Vectors document similarity measures to find documents in graylists similar to Pakistani blacklist documents. We would also compare the graylist documents of religious debates to the whitelist scripture references.

Table 1. Classified Corpus

Group	Nature	Reference URL
Pakistani blacklist	Anti-religious	http://propakistani.pk/wp-content/uploads/2010/05/blocked.html
Australian / Danish blacklists	Sexual	https://wikileaks.org/wiki/Leaked_Australian_blacklist_reveals_banned_sites https://wikileaks.org/wiki/Australian_government_secret_ACMA_internet_censorship_blacklist,_18_Mar_2009 https://wikileaks.org/wiki/Australian_government_secret_ACMA_internet_censorship_blacklist,_11_Mar_2009 https://wikileaks.org/wiki/Australian_government_secret_ACMA_internet_censorship_blacklist,_6_Aug_2008 https://wikileaks.org/wiki/Denmark:_3863_sites_on_censorship_list,_Feb_2008
Graylist	Religious and Social	Blogs, Debates, Reddit comments: eg. Alisina.org, anti-cair-net.org
Whitelist	Religious and Academic	KJV Bible CS course http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
Unrated	Newsgroup	http://qwone.com/~jason/20Newsgroups/
Banned wordlists	Sexual, anti-social, provocative	http://www.cs.cmu.edu/~biglou/resources/bad-words.txt http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/ https://gist.github.com/jamiew/1112488 http://bannedwordlist.com
Vocabulary	Newsgroup	http://qwone.com/~jason/20Newsgroups/vocabulary.txt

1. First we indexed the Australian/Danish content to see if term vectors accurately represented the banned words of sexual nature. We build our positional indexes in incremental fashion in a contextual window of 12 words. We searched the indexed corpora for the banned keywords list and noted the term vectors that matched. Then we indexed unrated content and checked if the blacklist content scored higher in term vector score for banned wordlist than the unrated content score, thus validating that learnt Semantic Vectors scores were reliable predictor of banned content.
2. Then we ran k-means clustering on the corpus to rediscover the context found within the corpus. We analyzed the nature of the content to see if any subtypes emerge from the banned content. Now we also indexed the Pakistani content of anti-religious nature. We also see if the banned content from sexual nature is separated from the anti-religious nature. The command for clustering is

```
java pitt.search.semanticvectors.ClusterResults
[banned_word_list]
```

3. We index the blacklists, graylists and whitelists and then run the cluster command on the entire vector store to find the documents that match the Pakistani blacklist document (of anti-religious content) and Australian and Danish blacklist document (of sexual content). This was purely unsupervised learning. No query terms were fed. We used the following command

```
java pitt.search.semanticvectors.ClusterVectorStore
      -numclusters 5 docvectors.bin
```

4. Then we indexed all corpora together and employed CompareTerms function to find the distance between blacklists of sexual nature from Australia and Denmark and anti-religious nature from Pakistan. Syntax of CompareTerms is

```
java pitt.search.semanticvectors.CompareTerms "<Table 2
Terms>" "<Table 5 Terms>"
```

Table 2. Term vectors found in Australian and Danish blacklist content that matched the four banned wordlists

dirty disease diseases disturbed dive doggystyle dong doom drug drunk drunken dumb dumbass dyke ejaculate ejaculated ejaculating ejaculation enema enemy erect erection ero escort ethnic european executed execution executioner explosion failed failure fairy faith fart farting fat fatass fear feces fellatio fetish fight filipina filipino fire firing fisting fondle fore foreskin foursome fuck fucka fuckable fucked fucker fuckers fuckin fucking fuckoff fucks fucktard funeral gangbang gangbanged gay gaysex geez german girls gob god goddamn goldenshower gross gun handjob harder harem headlights hebe hell henhouse hole homicide homo homosexual hook hooker hookers hooters horn horny husky hymen idiot illegal incest intercourse interracial israel itch jackoff jade jap japanese jeez jerkoff jesus jew jism jizz joint kid kill killed killer killing kink kinky knife knockers krap ky laid latin lesbian lesbo lez lezbo lezz liberal libido lies lingerie livesex lolita looser loser lotion lsd mad mams masterbate masturbate masturbating mexican milf molest moron motherfucker muff murder naked narcotic nasty nazi nigga nip nipple nude nymph oral orgasm orgies orgy panties peck pecker pee peehole penetration penis penises period perv piss pissed pissing pistol playboy poop pooper porn porno pornography pot poverty premature prick pros prostitute pubic puke pussies pussy pussylips quickie quim racist radical randy rape raped rapist rectum redlight remains retarded ribbed rimjob rimming robber satan scat schlong screw scrotum scum semen servant sex sexed sexpot sextoy sextoys sexual sexually sexy shag shagging shit shite shits shitty shoot shooting sick sissy skank slant slapper slaughter slave slime slut sluts slutty slutwear smack smut snatch sniggered snot sob sodomy spank sperm spit spreadeagle spunk spunky strapon stroke stroking stupid suck sucker swallow sweetness taboo tampon tang teat terror testicles threesome threeway tit titfuck titjob tits tittie titties titty toilet tongue torture tranny transexual transsexual transvestite twat twink uk upskirt urinate urine vagina vaginal vibrator violence virgin vulva wank wanking weapon whites whore willy womens wtf xxx

We also have to find representative query terms for graylists, whitelists and blacklists and compare them. For this we took the document topics as the query terms for comparison. We also validated our comparison technique in this step by comparing unrated content based on the document topic categories.

5. We took the most representative sample of blacklist and whitelist documents to search similar documents. The generic command for this was

```
java pitt.search.semanticvectors.Search -queryvectorfile
docvectors.bin -searchvectorfile termvectors.bin -
matchcase PATH_FOR_DOCUMENT_TO_MATCH
```


6. Then we filtered the blacklisted documents by applying the NOT keyword and visualize our vectors into normative tag clouds to see if Semantic Vectors could filter the blacklists from all corpora.

V. Results and Discussion

For step 1, 29651 terms and 13 documents were indexed. We found 315 term vectors that we indexed from Australian and Danish sexually provocative banned content to match up with banned words, the total supplied list of 1767 keywords. Among the term vectors matched in Australian and Danish banned content and banned keywords from Urbano's, CMU, JamieW and BannedWordList.com are Table 2:

We searched the indexed documents for the banned keywords. Among the similar terms searched by Semantic Vectors from the same corpus of Australian and Danish banned content, the following new words matched closely:

Table 3. Contextual words similar to the term vectors in Table 2 found by Semantic Vectors

0.999219:rich	0.998392:skinny
0.998824:creamy	0.998363:socks
0.998750:doll	0.998311:black
0.998685:trimmed	0.998296:natural
0.998610:tanned	0.998274:bus
0.998604:wanna	0.998222:total
0.998445:loving	0.998210:camera
0.998442:likes	0.998186:neighbor

These results show that the keywords are utilized 315/1767 = 17% in our corpus, which is a representative sample for the banned content.

We can use the same matched banned words in our query terms in our larger corpus set. The new words found by our semantic vectors encode the particular context in which the banned words are most likely to be used. For example, physical features like 'tanned', 'trimmed', 'creamy', 'skinny', 'black', and 'natural'; location related terms like 'neighbour' 'bus'; related accessories like 'camera' and 'socks', and actions like 'loving' and 'likes'.

Here we were using positional indexes that approached the indexing problem from the term-window perspective, keeping 12 words within the context of the term occurrence as highly weighted.

When searching the same banned wordlist (in Table 2) from the unrated content of the control experiment, we saw results significantly lower in match value with meaningless words like 0.983995:abcd, 0.983995:clio. This validated our hypothesis that the blacklists were represented by the learned term vectors in Table 2 and Table 3. Thus, we can use Semantic Vectors to compare graylists with whitelists and blacklists. But first, we have to get representative query terms for whitelists and graylists in order to compare.

In Step 2, on creating clusters of the new words, we found the clusters too close enough in contextual space to have any inter-cluster differences. However, cluster 4 is more male-oriented and Cluster 2 and 3 more female-oriented:

Table 4. General clusters of classified documents from Australian and Danish blacklist content

Cluster 1	Cluster 2	Cluster 3	Cluster 4
rich natural whore skinny	doll tanned socks	trimmed creamy wanna titties neighbor	banging escort

We incrementally added Pakistani dataset of anti-religious content. New 33319 term vectors and 9 doc vectors were added. Out of a total of 1821 banned words list, 325 sexually provocative and other offensive words were found in the corpus. Some new religiously motivated words not in the 4 banned word list were found , e.g. 'Rabbi', 'Fatah', 'Vatican', 'Usama', 'enemy'. They were pointing to some controversial usage, however, a few words were even sexual, making it double offensive and possibly 'blasphemous' in nature.

Table 5. Term vectors found from Pakistani blacklist content that matched with the four banned wordlists

deposit desire destroy devil dick dickhead dickweed die died dies dildo dingleberry dipshit dirty disease diseases disturbed dive doom dope drug drunk drunken dumb dumbass ejaculation enemy erect erection ethnic european excrement execute executed execution executioner explosion fag faggot failed failure fairy faith fart fat fatah fear feces fight filipino fire firing fore fraud fu fubar fuck fucked fucker fuckers fuckin fucking fucks fuk funeral gangbang gangsta gay geez genital german gin girls god goyim gross gun hamas harder harem hell henhouse heterosexual hijack hijacker hijacking hitler hitlerism hiv ho hoes hole homicide homo homosexual hook horn horny horseshit hostage hymen idiot illegal incest intercourse interracial israel israeli israels italiano jackass jade japanese jesus jew jewish jihad jizz joint kaffir kafir kid kill killed killer killing kills kink kkk knife ky laid latin lesbian liberal licker lies liquor loser lowlife lsd lucifer lynch mad mafia marijuana masterbate masturbating meth mexican mideast milf minority moles molest molestation molester mormon moron moslem motherfucker muff murder murderer muslim naked nasty nazi negro negroes niger nigerian nigerians niggardly nigger nude nuke oral orgasm orgy osama paki palestinian pansy panties peck pee penetration penile penis penises period perv pimp piss pissed pisses pissing pistol playboy pommy poo poop pooper porn pornography pot poverty premature prick propaganda prostitute protestant pubic puke pussies pussy rabbi racial racist radical radicals raghead randy rape raped raper rapist rectum redneck refugee reject remains republican retard retarded roach robber satan screw scum servant sex sexual sexually sexy shagging shit shite shithead shits shitty shoot shooting sick sissy slant slaughter slave slime slut sluts slutty smack snatch sniper snot sob sodom sodomy sonofabitch sos soviet sperm spit stringer stroke stupid suck sucker suicide swallow swastika sweetness taboo tarbaby terror terrorist teste testicles threesome tit tits toilet tongue torture towelhead transvestite trojan twat uk urinate urine usama vagina vatican violence virgin vomit weapon weewee welfare whiskey whites whitey whiz whore whorehouse willie wn womens wtf wuss yankee

For running search of banned word lists inside the indexed content with search type MINSIM meaning farthest conjunction and MAXSIM meaning closest conjunction, the new terms found were quite the opposite of what we would expect: they appeared to have positive sentiments, not provocative at all. Hence, Semantic Vectors found the irony here in usage of overtly positive words to represent the context of the found banned word list for anti-religious content. This also meant that in anti-religious content, offensive words are silver-plated by positive context.

Table 6. Contextual keywords opposite to the polemical term vectors in Table 5 found by Semantic Vectors

0.990560:love	0.980071:plus
0.990046:went	0.979457:game
0.989750:dark	0.978914:names
0.987048:putting	0.978677:white
0.984975:excuse	0.978452:your
0.984267:red	0.978309:read
0.983945:hearing	0.977609:fingers
0.980677:share	0.976367:join

The word clusters found out of the indexed Pakistani banned content contained some argumentatively emphatic keywords. Cluster 1 and 2 showed disapproval while Cluster 3 posed strongly affirmative. However, none of the banned words of sexual and provocative content came up as general sense out of the context specified in clusters below.

Table 7. General clusters of classified documents from Pakistani blacklist content of anti-religious nature

Cluster 1	Cluster 2	Cluster 3	Cluster 4
hard beyond excuse	how little time	went read sure real	putting hearing

Just for confirmation from the domain, the nine documents in the Pakistani banned content are from the websites:

Table 8. Website names of the Pakistani banned content

Bare Naked Islam	1.4 MB
Denmark vs Mohammed	915 kB
Draw Muhammad Day	248 kB
Jihad Watch Joseph Zaalishvilli	218 kB
Jihad Watch Nicolai Sennels	535 kB
Jihad Watch Raymond Ibrahim	203 kB
Jihad Watch Rebecca Bynum	549 kB
Jihad Watch Robert Spencer	478 kB

A cursory reading from the texts show the polemical keywords found in the text were not represented in the summary clusters. Therefore we selected the religiously motivated words from the list in Table 5 and rerun the Search and ClusterResults functions.

The results of both functions brought just one cohesive cluster of the following terms defining the entire nine document Pakistani banned content. The one word 'Christian' was a representative word showing all the websites in table 8 had that perspective in common.

Table 9. Single cluster representing websites in Table 8 from the Pakistani blacklist

Cluster 1
however leave government christian back well similar rest cases complete gone taking good please those glad

In Step 3 we added the whitelist, graylist and blacklist and created clusters. The following emerged:

Table 10. Automatic clusters of keywords from all whitelists, graylists and blacklists taken together

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
catholic	destroyed	unnecessary influential	christian murder soldiers	war education cause
Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
support basis	jew	killed wherever	general disagree	destroy direct country

Table 11. Automatic clustering of documents

Cluster 0	AllCorpus\Bare Naked Islam AllCorpus\Draw Muhammad Day AllCorpus\anti-cair AllCorpus\alisina.org AllCorpus\Denmark vs Mohammed AllCorpus\Jihad Watch Nicolai Sennels AllCorpus\Jihad Watch Rebecca Bynum AllCorpus\infidels AllCorpus\Jihad Watch Joseph Zaalishvilli AllCorpus\answering muslims AllCorpus\Jihad Watch Robert Spencer AllCorpus\Jihad Watch Raymond Ibrahim AllCorpus\arabs for israel AllCorpus\mistakes in quran AllCorpus\answering islam AllCorpus\al-rasooli **AllCorpus\Denmark URL List 2008** **AllCorpus\cutie art modeling studios**	
Cluster 1	AllCorpus\Australia URL list march 2009	
Cluster 2	AllCorpus\John_Chapter (1) AllCorpus\Matthew Chapter (7) AllCorpus\Matthew Chapter (14) AllCorpus\John_Chapter (7) AllCorpus\Luke Chapter (4) AllCorpus\Luke Chapter (23) AllCorpus\Mark Chapter (2) AllCorpus\Mark Chapter (8) AllCorpus\Luke Chapter (13) AllCorpus\John_Chapter (15) AllCorpus\Luke Chapter (19) AllCorpus\Matthew Chapter (23) AllCorpus\Mark Chapter (11) AllCorpus\Matthew Chapter (6) AllCorpus\Matthew Chapter (13) AllCorpus\John_Chapter (6) AllCorpus\Matthew Chapter (19) AllCorpus\Luke Chapter (3) AllCorpus\Luke Chapter (22) AllCorpus\Mark Chapter (1) AllCorpus\Luke Chapter (9) AllCorpus\Mark Chapter (7) AllCorpus\Luke Chapter (12) AllCorpus\John_Chapter (14) AllCorpus\Luke Chapter (18) AllCorpus\Matthew Chapter (22) AllCorpus\Mark Chapter (10) AllCorpus\Matthew Chapter (5) AllCorpus\Matthew Chapter (28) AllCorpus\Matthew Chapter (12) AllCorpus\Mark Chapter (16) AllCorpus\John_Chapter (5) AllCorpus\Matthew Chapter (18) AllCorpus\Luke Chapter (2) AllCorpus\Luke Chapter (21) AllCorpus\Luke Chapter (8) AllCorpus\Mark Chapter (6) AllCorpus\Luke Chapter (11) AllCorpus\John_Chapter (13) AllCorpus\Luke Chapter (17) AllCorpus\Luke Chapter (1) AllCorpus\John_Chapter (19) AllCorpus\Matthew Chapter (21) AllCorpus\Matthew Chapter (4) AllCorpus\Matthew Chapter (27)	AllCorpus\Matthew Chapter (11) AllCorpus\Mark Chapter (15) AllCorpus\John_Chapter (4) AllCorpus\Matthew Chapter (17) AllCorpus\Luke Chapter (20) AllCorpus\Luke Chapter (7) AllCorpus\Mark Chapter (5) AllCorpus\Luke Chapter (10) AllCorpus\John_Chapter (12) AllCorpus\Luke Chapter (16) AllCorpus\John_Chapter (18) AllCorpus\Matthew Chapter (20) AllCorpus\Matthew Chapter (3) AllCorpus\Matthew Chapter (26) AllCorpus\Matthew Chapter (10) AllCorpus\Mark Chapter (14) AllCorpus\John_Chapter (3) AllCorpus\Matthew Chapter (9) AllCorpus\Matthew Chapter (16) AllCorpus\John_Chapter (9) AllCorpus\John_Chapter (21) AllCorpus\Luke Chapter (6) AllCorpus\Mark Chapter (4) AllCorpus\John_Chapter (11) AllCorpus\Luke Chapter (15) AllCorpus\John_Chapter (17) AllCorpus\Matthew Chapter (2) AllCorpus\Matthew Chapter (25) AllCorpus\Mark Chapter (13) **AllCorpus\xnxx sex stories** AllCorpus\John_Chapter (2) AllCorpus\Matthew Chapter (8) AllCorpus\Matthew Chapter (15) AllCorpus\John_Chapter (8) AllCorpus\John_Chapter (20) AllCorpus\Luke Chapter (5) AllCorpus\Mark Chapter (3) AllCorpus\John_Chapter (10) AllCorpus\Mark Chapter (9) AllCorpus\Luke Chapter (14) AllCorpus\John_Chapter (16) AllCorpus\Luke Chapter (24) AllCorpus\Matthew Chapter (1) AllCorpus\Matthew Chapter (24) AllCorpus\Mark Chapter (12)
Cluster 3	AllCorpus\street meat asia AllCorpus\teenport	
Cluster 4	AllCorpus\girls no nude **AllCorpus\reddit comments.csv** AllCorpus\moped dolls AllCorpus\Adult friend finder AllCorpus\barbarian movies **AllCorpus\course** AllCorpus\pt reality AllCorpus\stream vid AllCorpus\non nude models AllCorpus\hardlyfucked	

When running document clusters function ClusterVectorStore, the results created 5 clusters from 121 docvectors, misplacing only five documents (shown in bold in Table 11). Cluster 0 is made up of anti-religious polemics banned by Pakistan. Cluster 1 is Australian list of banned URLs that is officially available in wikileaks. Cluster 2 is whitelist of scriptures. Cluster 4 and 5 are sexual content banned in Australia and Denmark. This makes the combined accuracy 116/121 = 95.9% for unsupervised learning.

In Step 4, comparing the Pakistani and Australian/Danish banned word list, the results showed the comparison ratio 0.903 , meaning the two word banned wordlists were 90% similar. Comparing unrated content, the documents labeled "religious.christian" and "atheism" showed comparison ratio 0.361 meaning only 36.1% matched between the two concepts, which validated our hypothesis for comparing whitelists and blacklists from graylists.

Comparing all the whitelist documents topics from graylists, 82.9% match occurred. Comparing all graylist documents topics to those of Pakistani blacklists and Australian/Danish blacklist documents showed percentage similarities of 83.0% and 82.7% respectively. This meant that the graylist documents full of controversial debates were more connected to the whitelist scriptures and blacklist anti-religious content and less to the sexually provocative content typically in Australian and Danish banned list. The score percentage is to be taken relative and not absolute, since the topics were manually labeled and most not found in vector list. Term comparison of banned wordlist showed more representative percentage (90.3%) of term vectors found and matched. The results of Step 4 are summarized in Figure 1.

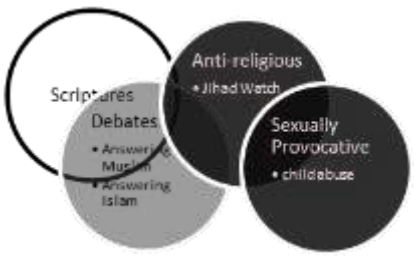

Fig. 1. Comparison of Whitelist containing Old Testament and New Testament, Graylist of polemic debates and Blacklist of anti-religious and sexually provocative websites

From Step 5 we learnt that finding the document relevance was more accurate starting from whitelist documents rather than looking for blacklist documents. For example, searching similar documents to "xnxx sex stories.txt" always returned scriptures incorrectly. However, searching for scriptures by a chapter of Luke or Matthew always returned the scriptures and nothing else. Similarly referring one graylist document to the rest of graylists and whitelists was more likely than referring blacklist document to graylist documents. This has benefit for automatic search queries, where searching for banned content most likely results in referring to acceptable content in whitelist rather than more banned content. This is because blacklist refer to rare itemset which is even less likely to be co-referred by another rare itemset.

Fig. 2. 2d plot of vectors show normative tag cloud of debate documents in graylist. Right boundary words of warriors, devil and perpetrator are stronger.

In order to visualize our filtration mechanism in Step 6, we first visualize normative tag cloud from graylist semantic vectors in Figure 2. For that we feed the topic list to PrincipalComponents function of the SemanticVectors.Viz package. The package reduces the vectors in 2d space and visualizes a limited number of summarizing word lists on to a tag cloud. Clusters in the tag cloud are easy to differentiate the senses similar and varying across the 2d space.

Fig. 3. 2d plot of vectors show normative tag cloud of graylist words after anti-religious word list is filtered. Cluster of the two words 'bombing' and 'affect' show some strong temper.

Figure 2 shows the keywords from the graylist topics related to debate polemics. Summary keywords like ‘devil’, ‘lies’, and ‘warriors’ refer to extreme polemics, while words supporting argumentation like ‘research’, ‘literature’, ‘faith’, ‘sacrifice’ and ‘support’ form the root and stem of the argument.

Fig. 4. 2d plot of vectors show normative tag cloud of graylist words after sexually provocative word list is completely filtered.

In Figure 3, the topics from banned Pakistani content are excluded. These topics were mainly anti-religious in context. After filtering these, the main keywords are spread to show matter-of-fact scenarios, like ‘attacks’, ‘sentenced’, ‘grief’, ‘question’ and ‘lack’. Some words form an insoluble lump like ‘bombing’ and ‘affect’ which show still a strong indication of argument.

After removing the main topics from the Australian and Denmark blacklists, the resulting Figure 4 finally shows a very politically correct set of words including ‘polite’, ‘unsuspecting’, ‘mission’, etc. Also, the given censored word set creates a well spread normative tag cloud of consistently appropriate keywords.

VI. Conclusion

What we observe and demonstrate from clusters and normative tag clouds generated using Semantic Vectors is that we can automatically filter unethical and objectionable content from the Internet using the technology. Semantic Vectors work best in unsupervised learning of document clusters, which can correctly place anti-religious, sexually explicit content out of the otherwise acceptable documents, by an accuracy of 95.9%. Semantic Vectors can also answer questions on the polarity of gray area documents, whether they are more similar to blacklists or whitelists. Search results can refer to more positive answers that remove the banned words and blacklist content. Another corollary of our study is that in order to grow the knowledge base of acceptable documents, a sample of whitelist documents can build similar whitelist document index better than a sample of blacklist documents to index similar others.

To note performance, a total of twenty thousand documents sizing 108 MB can emit 620603 term vectors and get indexed in only 172 seconds. Once indexed, querying and searching is faster and more accurate using small query set. Hence, Semantic Vectors are most suited for big data text analytics and ethical filtering over the cloud. Also, the Semantic Vectors package is language independent. We have used English language corpus with some foreign terms like rabbi and jihad which were recognized. However, future work may use Semantic Vectors in complete training and clustering of documents based on ethical features in Asian languages.

References

[1] Z. Nabi. “The Anatomy of Web Censorship in Pakistan”, arXiv:1307.1144v1 [cs.CY] (2013).

[2] A. Chaabane, T. Chen, M. Cunche, ED. Christofaro, A. Friedman, M.A. Kaafar. “Censorship in the Wild: Analyzing Internet Filtering in Syria”, arXiv:1402.3401v3 [cs.CY], (2014).

[3] E Veer. "Staring: how Facebook facilitates the breaking of social norms." *Research in Consumer Behavior* 13, 185-198, (2011).

[4] R.J. Diebert, J.G. Palfrey, R. Rohozinsky, J. Zittrain. *Access Denied: The Practice and Policy of Global Internet Filtering*, The MIT Press, ISBN-10:0-262-54196-3, ISBN-13:978-0-262-54196-1, (2008).

[5] R.J. Deibert, J.G. Palfrey, R. Rohozinsky, J. Zittrain. *Access Controlled: The Shaping of Power, Rights and Rule in Cyberspace*, The MIT Press. ISBN: 9780262514354, (2010)

[6] Internet Content Filtering and Blocking: Electronic Frontiers Australia, (2006)

[7] D. Widdows, T. Cohen. “The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics”, Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010), (2010).

[8] T. Bohne, S. Rönnau, and U.M. Borghoff. Efficient keyword extraction for meaningful document perception. In Proceedings of the 11th ACM symposium on Document engineering (DocEng '11). ACM, New York, NY, USA, 185-194. DOI=10.1145/2034691.2034732, (2011)

[9] L. Lessig. *Code and Other Laws of Cyberspace.* New York: Basic Books, (2006).

[10] J. Zittrain. *The Future of The Internet and How to Stop It.* Yale University Press. ISBN 978-0-300-15124-4, (2008)

[11] L.M. Shaikh, S. Sarfraz, A.N. Khan. “PsycheTagger: Using Hidden Markov Model to annotate English Text with semantic tags based on emotive content”. In Proceedings: AIKED’12 Proceedings of the 11th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Pages 219-224, World Scientific and Engineering Academy and Society (WSEAS), Steven Point, Wisconsin, USA, ISBN: 978-1-61804-068-8, (2012).

[12] A.N. Khan, M. Aslam, A.M. Enríquez. “Mining for Norms in Clouds: Complying to Ethical Communication through Cloud Text Data Mining”, in Proceedings of Fifth IEEE UCC, *IEEE Xplore*, Print ISBN: 978-1-4673-4432-6, (2012).

[13] K. Govinda, P.K. Abaru, G.P. Reddy. "On-Demand Secure Streaming of Multimedia Data over Cloud", *International Journal of Engineering and Technology*, Vol 5 No 3, ISSN: 0975-4024 (2013)
[14] T. Zhu, D. Phipps, A. Pridgen, J.R. Crandall, and D.S. Wallach. The velocity of censorship: high-fidelity detection of microblog post deletions. In *Proceedings of the 22nd USENIX conference on Security* (SEC'13). USENIX Association, Berkeley, CA, USA, 227-240, (2013).
[15] J. McLachlan and N. Hopper. "On the Risks of Serving Whenever You Surf: Vulnerabilities in Tor's Blocking Resistance Design". In WPES, (2009).
[16] M. Dusi, M. Crotti, F. Gringoli, and L. Salgarelli. "Tunnel Hunter: Detecting Application-layer Tunnels with Statistical Fingerprinting." Computer Networks, 53(1):81–97, (2009).
[17] C. Leberknight, M. Chiang, H. Poor, and F. Wong. "A Taxonomy of Internet Censorship and Anti-censorship", (2012).
[18] S.H. Jan. "Internet Filtering Software Tests", San José Public Library, Revised Report, (2008).
[19] H. P. Luhn. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 2, 159–165, (1958).
[20] G. Salton, C. Buckley. "Term-weighting approaches in automatic text retrieval", Inf. Process. Manage. 24, 5, 513–523, (1988).
[21] P. Song, A. Shu, A. Zhou, D. S. Wallach, J. R. Crandall. A pointillism approach for natural language processing of social media. In IEEE International Conference on Natural Language Processing and Knowledge Engineering, arXiv:1206.4958v1 [cs.IR], (2012).
[22] S. Nazirova. Anti-Spam Software for Detecting Information Attacks, I.J. Intelligent Systems and Applications, 10, 25-34, (2012).

Authors' Profiles

Ahsan N. Khan is a Data Scientist working with Teradata. This research work is the output of his Masters of Science Thesis in University of Engineering and Technology, Lahore. Mr. Khan is an active writer and blogger, with several research publications listed in Amazon, Google Scholar, Microsoft Academic Research and WSEAS. He has also graduated Magna cum Laude in the same field as Bachelors of Science from National University of Computer and Emerging Sciences. Hence his studies and experience in the data science is spanned over a decade of quality work contributions, including his books on theses, *Text Psyche Mining: Normative Vision* and *Normative Tag Cloud: Creating Normative Word Sense Spectrum*, published with Amazon.com.

How to cite this paper: Ahsan N. Khan,"Automatic Ethical Filtering using Semantic Vectors Creating Normative Tag Cloud from Big Data", International Journal of Intelligent Systems and Applications (IJISA), vol.7, no.4, pp.17-25, 2015. DOI: 10.5815/ijisa.2015.04.03

