Field-Theoretic Modeling Method for Emotional Context in Social Media: Theory and Case Study

Monte Hancock[1(✉)], Shakeel Rajwani[2], Chloe Lo[2], Suraj Sood[2], Elijah Kresses[2], Cheryl Bleasdale[2], Nathan Dunkel[2], Elise Do[2], Gareth Rees[2], Jared Steirs[2], Christopher Romero[2], Dan Strohschein[3], Keith Powell[2], Rob French[2], Nicholas Fedosenko[2], and Chris Casimir[2]

[1] 4Digital, National Aeronautics and Space Administration, Washington, DC, USA
mfhancock@aol.com
[2] Sirius15, National Aeronautics and Space Administration, Washington, DC, USA
[3] National Aeronautics and Space Administration, Washington, DC, USA

Abstract. Just as masses and charges give rise to gravitational and electric fields, the online behaviors of individuals engaged in online social discourse give rise to an "emotional context" that conditions, and is conditioned by, these behaviors. Using Information Geometry and Unsupervised Learning, we have formulated a mathematical field theory for modeling online emotional context. This theory has been used to create a soft-ware application, **Sirius15**, that infers, characterizes, and visualizes the field structure ("emotional context") arising from this discourse. A mathematical approach is presented to social media modeling that enables automated characterization and analysis of the emotional context associated with social media interactions. The results of a small, preliminary case study carried out by our team are presented.

1 Background

Human collaborative processes are being revolutionized by the emergence of ubiquitous, completely portable social media. Group decision-making, social interaction, educational instruction, and many other directed and undirected cognitive interactions are now conducted without a single spoken word being exchanged.

Augmented cognition is a form of human-systems interaction in which a tight coupling between user and computer is achieved via physiological and neurophysiological sensing of a user's cognitive state. An essential, sometimes determinative component of this state is the user's emotional state; this not only affects his or her cognition, but the cognition of others by means of the way individual online behaviors condition the emotional context within which interactions take place.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 418–425, 2015.
DOI: 10.1007/978-3-319-20816-9_40

The underlying motivating principles are:

1. By the term "emotional context" we mean a vector field generated by documents posted on a specific social medium. The documents are themselves within this attribute space.
2. Text sources (i.e., persons) can be regarded as short-memory text generators having goals and a partially-observable history.
3. The output text (forum posts on social media) is purposeful and sequential in time.
4. The output text is constrained by the medium and the medium's linguistics.
5. The text is generated within an "emotional context" that both informs, and is informed by, each generator.

We take 4 above as our guiding principle, since all field theories are ultimately based upon the idea that,

"The entities whose actions are conditioned by a field are themselves its sources and sustainers."

We begin by assuming that this is just as true of the emotion that drives social behaviors as it is of the massive objects that drive gravitational behaviors. The field can then be viewed as a mathematical device for explaining how entities influence one another without directly interacting.

2 The Data

Sirius15 has been benchmarked using two years of colloquial, natural-language discourse from an English-language blog site. No prior domain assumptions are made (e.g., the method is language independent and does not require "translation" in order to operate). Results from the benchmark indicate that content-clustering is supported, and that "social signatures" of individual posters can be characterized.

The figure below shows a short sequence of blog posts from the data used to create the benchmark.

Post #	Thread #	Post_in_thread	Poster ID	Post
85527	4344	18	297	Damn it Small, how the hell do you get tackled by the Punter! ! !? !? !? !
85528	4344	19	377	That's alright... He's smaller than the punter!
85529	4344	20	297	Well, 3rd and 11... let's see what they do...
85562	4345	1	0	Ricky Dobbs back at it Out the past 2 weeks and Navy felt it. Winning against Wake 13-10 and l
85563	4345	2	181	Sorry Vespula, take out Nesbitt and put Ricky Dobbs in GA Tech's option. forget about it.
85564	4345	3	13	Uh oh, Navy up 14-0
85566	4345	5	297	GO NAVY! !
85567	4345	6	13	Now he throws a 50+ yard TD. Dobbs for Heisman.
85568	4345	7	297	GO NAVY GO!
85569	4345	8	372	Where is Bks for this?
85570	4345	9	6	Go Midshipmen! !
85572	4345	11	128	GO NAVY! !
85574	4345	13	13	Yes, but he isn't Ricky Dobbs.
85575	4345	14	415	and Navy wins! ! ! ! ! ! !
85576	4345	15	20	Uhh I wasn't aware that you could onside kick after a safety. I thought you had to punt/kick it de
85577	4345	16	20	So Navy gets the safety to make it 23-14 and Notre Dame onside kicks the safety kick? ? and rec
85578	4345	17	128	YEAH,BABY! GO NAVY!
85579	4345	18	20	Navy recovers the onside kick and runs out the clock. Game = Navy2 times in 3 years for the Mi
85580	4346	1	0	OH MY GOSH MY HEAD IS GOING TO EXPLODE I hate living in Oklahoma. The newspaper says
85581	4346	2	111	Get DirectTV. lots of options there...
85582	4346	3	377	'Stro, I'm having the same problem. Thank goodness Comcast settled finally with ESPN360 and
85583	4346	4	420	try justin tv. . I love that site now. . NO HD quality but It's better than nothing. .
85584	4346	5	181	TG for AT&T, they are an affiliate of ESPN360.
85585	4346	6	297	Watching it on ABC with me Mum sorry Stro

3 Approach

The ultimate goal of this work is to extend and mature emotion-mining applications that will inform practical action, either in the real-world, or in the social medium itself.

While much work has been done (and remains to be done) in machine translation of colloquial text, relatively little formal mathematical analysis of the ambient emotional context that emerges from the interaction of humans using social media has been published. The **Sirius15** application ingests bulk social media data and, by means of a six-phase procedure, infers an empirical vector field structure characterizing the emotional attributes of the discourse. This we call the media's "Emotion Field".

The Six Phase Processing Sequence

Phase_0_exe
(Ingest and conform raw forum post set, create initial word list)
input: Forum.csv
output: Phase_I_In.csv

Phase_I.exe
(Ingests original word list with repetitions, annotates with idf scores)
input: Phase_I_In.csv
output: Phase_I_Out.csv
Wordstring_I.txt

Phase_II.exe
(filter out misspelled words, renumber threads in annotated word list and forum text)
input: validwords.csv
Phase_I_Out.csv
Forum.txt
output: Phase_II_Out.csv
Forum_II.txt

Phase_III.exe
(Create idf-weighted angles-only distance matrix, check metric consistency)
input: Phase_II_out.csv
validwords.csv
output: Phase_III_out.csv
dmFAIL.csv
trifail.csv

Phase_IV.exe
(Generate a Torgerson Coordinate feature vector file having selected dimension)
input: Phase_III_out.csv
output: Phase_IV_out.csv
Rowrand_IV.csv

Phase_V.exe
(Cluster the Torgerson Coordinate feature vector file, produce C45 file and summaries)
input: Phase_IV_out.csv
output: Phase_V.hOT
Phase_V_cl.csv
P5_"clus#".csv (if selected; could be up to 999 files)

Modern text mining methods generally rely on a combination of statistical and graph-theoretic schemas for representing information. These schemas are parsed and quantified to obtain information about associations, processes, and conditional probabilities for variables of interest. While some automation exists, the semantic characterization of corpora is largely manual and ad hoc. The state of the art is described in detail in [1].

The shortcomings of current approaches arise largely because each was designed fairly specifically to overcome failures observed in others. Because a "trouble-shooting" mentality is inherently ad hoc, none of the current methods is founded upon a mathematical formalism intended to cover the entire problem space. Each of these proposed solutions has given rise to new problems. Our work re-addresses this space by employing a broader and more mature mathematical foundation.

The six-phase **Sirius15** processing sequence begins by fusing a collection of document metrics to create a matrix of pair wise distances between social media posts. This can be done in a coordinate-free way, since many document metrics are statistical rather than "vectorial" in nature (e.g., Tf.idf, term histograms). The combined differences between the selected metrics are taken as a similarity measure, where greater differences imply a greater distance between the underlying documents.

From this document distance matrix is inferred a set of points in an N-dimensional Euclidean space, with each point representing a single document in the corpus. The points are developed as a set to have the same distance matrix as the corpus of documents. This embedding geometrizes the document analysis problem, facilitating the use of many mature data mining tools that require numerical (rather than nominal) input features.

The figures immediately below are snippets collected during the model building process:

Bag-Of-Words for Thread 4345

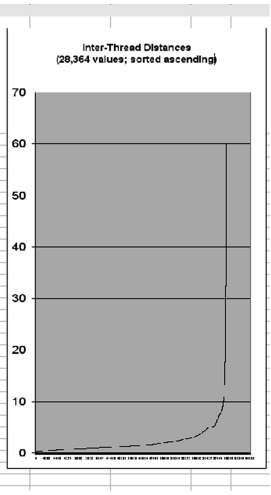
[RICKY] [DOBBS] [PAST] [2] [WEEKS] [NAVY] [WINNING] [WAKE] [LOSING] [GUESS] [SCORED] [SEASON] [QUARTER] [NOTRE] [KID] [RICKY] [DOBBS] [FORGET] [UH] [NAVY] [NAVY] [DAME] [OUTSCORE] [STOP] [THROWS] [YARD] [DOBBS] [NAVY] [BKS] [NAVY] [NAVY] [PAST] [2] [WEEKS] [NAVY] [WINNING] [WAKE] [LOSING] [GUESS] [SCORED] [SEASON] [QUARTER] [NOTRE] [KID] [BACKUP] [PROCTOR] [DECENT] [MENTION] [DAUGHTER] [KID] [RICKY] [NAVY] [AWARE] [KICK] [NAVY] [SAFETY] [NOTRE] [DAME] [KICKS] [SAFETY] [RECOVERS] [DRIVE] [NAVY] [SECS] [KICK] [NAVY] [RECOVERS] [KICK] [RUNS] [GAME] [3] [CONGRATS] [NAVY] [WINNING] [RECORD] [RECORD] [WINNING] [TEAMS] [RIDICULOUS] [RUN] [BKS] [CROW] [BOARD]

```
4344,SAFETY, 3.690218
4344,SAFETY, 3.690218
4344,RECORD, 2.432366
4344,RECORD, 2.432366
4345,OKLAHOMA, 2.759742
4345,GEORGIA, 2.489823
4346,SOUTHERN, 3.095511
4346,ILLINOIS, 3.384836
4346,VILLANOVA, 4.882356
4346,RICHMOND, 4.797799
4346,RICHMOND, 4.797799
4346,RICHMOND, 4.797799
4347,OREGON, 2.745756
4347,ARIZONA, 3.418284
4347,STANFORD, 3.793596
4347,STANFORD, 3.793596
4347,THREAD, 1.975777
```

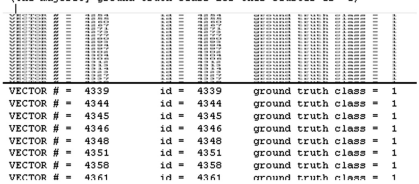
idf (inverse document frequency) scores indicate the corpus-wide significance of individual terms.

The snippet depicted here is from an experiment where only terms having at least six symbols were allowed.

Post #	Thread A	Thread B	Distance
1	1	2	0.978551603
2	1	3	1.086500977
3	1	4	1.505982586
4	1	5	1.098746197
5	1	6	1.103827229
6	1	7	2.951201197
7	1	8	5.011412709
8	1	9	0.982489654
9	1	10	1.659586755
10	1	11	1.055342659
11	1	12	2.194336924
12	1	13	1.084145534
13	1	14	1.339357735
14	1	15	1.333951271
15	1	16	0.995047952
16	1	17	1.05945775
17	1	18	1.099512776
18	1	19	1.412301333
19	1	20	1.138410636
20	1	21	0.988608828
21	1	22	1.495154694
22	1	23	1.601018942
23	1	24	1.206200763
24	1	25	0.9846962
25	1	26	1.033819348
26	1	27	1.127899311
27	1	28	1.04706031
28	1	29	1.223836906
29	1	30	1.003666911
30	1	31	1.28506844
31	1	32	1.098203747
32	1	33	1.074013292
33	1	34	1.325868347
34	1	35	1.162138896
35	1	36	0.99543964
36	1	37	1.047864704
37	1	38	1.1628761
38	1	39	1.156593155
39	1	40	1.842892939
40	1	41	1.926592621




```
VECTOR # =  2401    id =  2401    ground truth class =  1
VECTOR # =  2406    id =  2406    ground truth class =  1
VECTOR # =  2412    id =  2412    ground truth class =  1
-------------------------------------------------------
VECTORS ASSIGNED TO CLUSTER  316 BY THE MACHINE:
(The majority ground truth class for this cluster is  1)
```



```
VECTOR # =  4339    id =  4339    ground truth class =  1
VECTOR # =  4344    id =  4344    ground truth class =  1
VECTOR # =  4345    id =  4345    ground truth class =  1
VECTOR # =  4346    id =  4346    ground truth class =  1
VECTOR # =  4348    id =  4348    ground truth class =  1
VECTOR # =  4351    id =  4351    ground truth class =  1
VECTOR # =  4358    id =  4358    ground truth class =  1
VECTOR # =  4361    id =  4361    ground truth class =  1
```

The distances derived between "threads" in the social medium were used to infer Euclidean coordinates for for threads. The resulting vectors were then clustered to determine Blog "Clique Structure".

All Phases of the application run under Microsoft Windows on a 1-core processor. The **Sirius15** prototype operates in a "batch" mode; all posts are assumed to be present when modeling is performed.

During model construction, computational complexity is $O(N^2)$ in the number of threads (a thread is a topically heterogeneous collection of posts). Our benchmark data set contained 4,487 threads. The experiment reported below is based upon a subset of 136 threads, consisting of a total of 2,847 posts.

A large-scale processing-complexity experiment had the following results:

Torgerson Coordinates inferred using the Laplacian of the Field Potential

Thread #	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Cluster #
1	1.132851	-0.7324935	-2.807379	-0.3270132	-1.127138	0.4688708	1
2	1.713233	-0.7147069	-3.051881	-0.6393204	-1.826719	0.4107876	17
3	0.7090547	-1.19348	-2.330333	-0.6497642	-1.06594	0.7433085	1
4	1.510779	-1.410487	-2.479628	-0.501007	-1.247655	0.603342	1
5	1.507237	-1.195071	-2.303251	-0.1950196	-1.16112	0.7608344	1
6	1.328545	-1.361132	-2.976217	-0.4525052	-1.259198	0.413531	1
7	1.234295	-1.261456	-2.197749	-0.547074	-1.170727	4.41E-02	1
8	-0.3268293	0.5176432	4.71E-02	0.9533091	0.1505481	1.71E-02	32
9	-0.8928862	0.7833325	-0.3112981	0.4345004	0.3846253	-0.1534715	32
10	-1.340502	-0.2307369	-0.4626657	-1.561791	1.752655	0.7273825	67
11	-1.590019	-0.3873197	-0.6359065	-1.117083	1.774765	0.8708808	67
12	-1.511013	-0.6160021	-0.4751602	-1.668562	1.521265	0.3149354	67
13	-1.366065	0.12779	-0.1892876	-1.577721	1.500369	0.5746704	67
14	-1.6567	-0.4571404	-0.5007457	-1.743836	1.283633	0.8670825	67
15	-1.308586	-3.09E-02	-0.2335659	-1.30887	1.754491	0.34057	67
16	-1.59445	0.1800286	-1.094328	-1.371261	2.15339	0.5787393	67
17	-1.674331	-0.3219829	-0.6503383	-1.427543	1.829624	1.084335	67
18	0.4031281	1.543705	-0.8230213	-1.761889	1.7895	1.277841	49
19	0.2925319	1.633054	-0.294793	-1.6646	2.074924	1.346127	49

Prototype Performance

Corpus: ~10 million words
Duration: 2 years
Posts: 87,983
Threads: 4,487
Members: 224
Machine: Intel i7 920 @ 2.67 GHz CPU (Windows XP)

Function Execution times

Phase_0: 6 seconds
Phase_I.exe: 22 seconds
Phase_II.exe: 1 hour and 28 minutes
Phase_III.exe: 2 hours, 59 minutes, and 13 seconds
Phase_IV.exe: 12 hours, 35 minutes, and 53 seconds (*)
Phase_IV_NCC.exe: 13 minutes, and 1 second
Phase_V.exe: 27 seconds

Tfidf__II.exe: 2 minutes and 29 seconds

(*) Operator discretion. Time is $O(n^2)$ in number of threads.

4 The Mathematics of the Approach

The field-theoretic approach gives a unifying mathematical framework for applications of computational linguistics to emotion mining akin to the framework Maxwell's Equations provide for Electromagnetism: a set of field equations that provide a rigorous mathematical framework for automating aspects of currently man-intensive characterization of the "emotional context" of online social behavior. The mathematics is relatively straightforward:

Let $\mathcal{F}$ be a collection of finite length character strings ("threads"):

$$\mathcal{F} = \{A_j\} = \{A_1, A_2, \cdots, A_M\}$$

Let $d_{ij}(A_i, A_j) = d_{ij}$ be a metric on $\mathcal{F}$. Form the distance matrix:

$$D(A_i, A_j) = [d_{ij}], \qquad \text{i,j} = 1, 2, \ldots \text{M}$$

This matrix will be symmetric, zero diagonal, and non − negative.

Let: $$S = \{\overrightarrow{a_j}\} = \{\overrightarrow{a_1}, \overrightarrow{a_2}, \cdots, \overrightarrow{a_M}\} \in \mathbb{R}^N$$

be a (hypothetical) set of vectors having distance matrix D. Regarding the $\overrightarrow{a_j}$ as field sources, we define a discrete scalar potential $\wp$ on S by:

$$\wp(\overrightarrow{a_i}) = g \sum_{j=1}^{M} \left(\left\| \overrightarrow{a_i} - \vec{a}_j \right\| - d_{ij} \right)^2 \quad (*)$$

Here g is a non-negative constant that can be chosen arbitrarily to facilitate specific anticipated applications of the theory (e.g., sensitivity analysis).

In general, the existence of such an S for a given D is not guaranteed; in particular, D might not exist for N=1, but exist for N=2. Also, because S is informed only by the distances between the $\vec{a}_j$, any rigid placement of a solution is also a solution. Therefore, a solution can be registered in $\mathbb{R}^N$ for convenience.

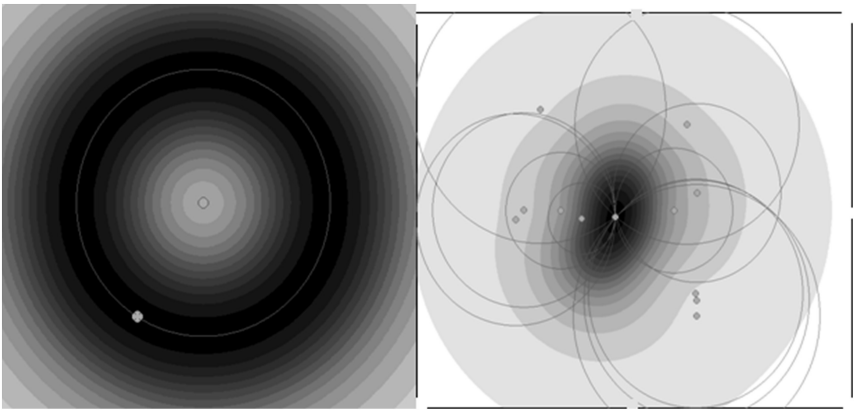
An approximate solution S for (*) can be found by various methods (e.g., SVD, Gradient Descent, Monte-Carlo).

It is laborious, but neither difficult nor particularly instructive to show that the formal gradient:

$$\nabla \wp(\overrightarrow{a_i}) = \mathrm{C}(\overrightarrow{a_i})$$

Is a vector field on S, and that S is precisely the set of zeros of the Laplacian of $\wp$ (which is to be expected).

The field equations give rise to a radially symmetric scalar potential, depicted in the figure on the left below. The field source is the center, which is a blog post that establishes an "energy well" that can be occupied by another document (at "7 o'clock").



The field satisfies the superposition principle, so as additional blog posts are generated, they can just be directly added in (the figure on the right above). Further, by the methods of Differential Geometry, it can be shown that the "emotion" field is conservative. In particular, it is path-independent, which implies that we need not retain the history of a post to understand its immediate effect on the emotion field. This "statelessness" is potentially important for future work, though we have not exploited it.

The derived attribute space is a Hilbert Space of appropriate dimension that creates coordinates in a natural way using unsupervised machine learning. The method was discovered independently by our team, but was first described by [2].

The dimension of the Euclidean Space can be chosen at will (though poor choices cause convergence problems; see below). We use a variation of the Delta Rule from machine-learning for inferring the embedding.

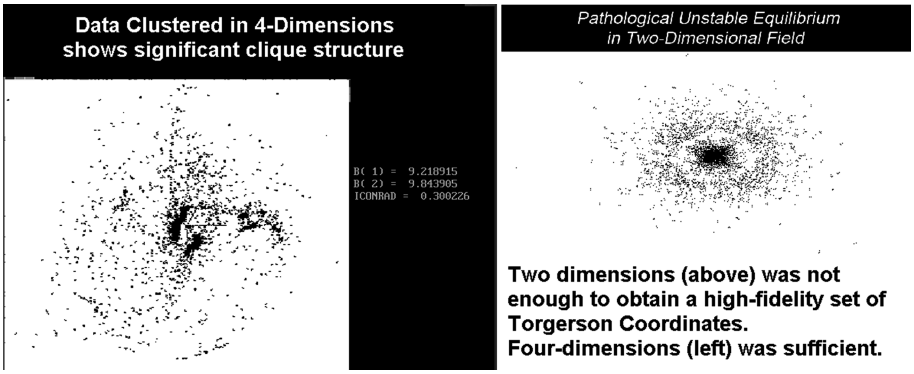
The Delta Rule is a first-order gradient descent method. When it is written as a distance minimization expression, it can be interpreted as a differential equation describing a vector field; a solution (to coordinatize the document data) is then a set of Lagrangian Points for this differential equation. In this way, the field is an emergent property of the points it positions, and the positions are constrained by the field. Specifics are given below.

Other methods can be used to coordinatize the data from the distance matrix (e.g., Singular Value Decomposition, Lagrange Multipliers, Newton's Method). We have found that convergence by gradient descent is usually not complete, owing largely to the fact that natural text metrics might not result in a fused distance function that is a metric (e.g., Triangle Inequality is satisfied only approximately). This is sometimes overcome by increasing the dimension of the Euclidean Space. We also use an adaptive convergence rate and an annealing schedule to speed up convergence.

Once the data are geometrized, data mining methods for data visualization, signature extraction, clustering, building classifiers, etc., can be used to complete the emotional context modeling process [3].

5 Future Work

The field-equations can be used to impose an inherent, a priori clustering of entities in the space. This clustering determines, and is determined by, the terrain geometry; these are in dynamic tension. Clusters, or cliques, correspond to "emotion plateaus" in the original feature space.



We have performed some preliminary work on the development of predictive analytics. Perhaps more interesting is the "domain segmentation" clustering provides, which might be used to identify those members of a forum most likely to engage in displays of emotive language. This is a type of Signaturing; done across the entire problem space, it constitutes a type of "Emotion Terrain-Forming."

In looking at the details of the benchmarked data set, the techniques discussed above are able to identify cliques of posters. More importantly, the "emotional distance" between cliques might be quantified, supporting assessment of the "emotional separation" of cliques, and individuals within cliques. Further, the level of emotional impact of certain terms might be numerically estimated.

References

1. Lee, S., Song, J., Kim, Y.: An empirical comparison of four text mining methods (2010)
2. Torgerson, Warren S.: Theory & Methods of Scaling. Wiley, New York (1958). ISBN 0-89874-722-8
3. Hancock, M.: Practical Data Mining. CRC Press, Boca Raton (2011). ISBN-13: 978-1439868362