UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

FONDO DI ATENEO PER LA RICERCA ANNO 2016
PROGETTO DI RICERCA DI DIPARTIMENTO

1.Titolo del Progetto di Ricerca: Creazione di una postazione mobile per l'analisi quantitativa di dati testuali

2. MacroSettore ERC del progetto: PE1

Sottosettori ERC di riferimento: PE1_14 (Statistics); PE1_18 (Scientific computing and data processing); PE1_21 (Application of mathematics in industry and society)

3.Parole Chiave (MASSIMO 5): Text analysis; Text mining;

4.Responsabile Progetto (P.I.) (ricercatore a tempo indeterminato e ricercatore a tempo determinato ex art. 24 L.240/2010, lettera a) e lettera b), professore associato o professore ordinario)

COGNOME: Martini

NOME: Maria Cristiana

Data di nascita: 20/08/71

Qualifica: Professore associato

Dipartimento: Dipartimento di Comunicazione ed Economia

(telefono): 0522-523232

(E-mail): cmartini@unimore.it

5.Sottosettore ERC del PI: PE1_14 (Statistics)

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

6.Elenco dei docenti e dei ricercatori (strutturati e non strutturati) partecipanti, appartenenti esclusivamente all'Università di Modena e Reggio Emilia

NOME	COGNOME	DIPARTIMENTO	RUOLO/ Tipologia di rapporto
Maria Cristiana	Martini	Comunicazione ed Economia	Professore associato

7. Principali pubblicazioni del P.I. e dei componenti il gruppo di ricerca negli ultimi 5 anni 2012-2016 (max 10), con eventuali indici bibliometrici:

Martini, Maria Cristiana; Fabbris, Luigi (2015) - *Beyond Employment Rate: A Multidimensional Indicator of Higher Education Effectiveness* - SOCIAL INDICATORS RESEARCH - pp. da 1 a 20 ISSN: 0303-8300

Martini, Maria Cristiana; Galli, Giovanna; Arseni, Chiara (2016) - *Brand Trust: un costrutto multidimensionale ed evolutivo* - MICRO & MACRO MARKETING - n. volume 1/2016 - pp. da 17 a 38 ISSN: 1121-4228

Maria Cristiana Martini; Cristiano Vanin (2013) - *A Measure of Poverty Based on the Rasch Model* (Torelli Nicola; Pesarin Fortunato; Bar-Hen Avner - *Advances in Theoretical and Applied Statistics*) (Springer-Verlag Berlin Heidelberg DEU) - Studies in Theoretical and Applied Statistics - pp. da 327 a 337 ISBN: 9783642355875 | 9783642355882

Fabbris L.; Martini M.C. (2013) - *Dimensions of graduates' job satisfaction in the short and medium terms* - STATISTICA APPLICATA - n. volume 23(1) - pp. da 9 a 36 ISSN: 1125-1964

Maria Cristiana, Martini (2013) - *Educational investments to guarantee graduates' human capital effectiveness* (Advances in Latent Variables - Methods, Models and Applications - Brescia - 19-21 june 2013) (Brentari E., Carpita M. - *Advances in Latent Variables*) (Vita e Pensiero Milano ITA) - pp. da 1 a 7 ISBN: 9788834325568

Martini, Maria Cristiana (2013) - *Visualisation and Analysis of Affiliation Networks as Tools to Describe Professional Profiles* (Giusti A., Ritter G., Vichi M. - *Classification and Data Mining*) (Springer-Verlag Berlin Heidelberg DEU) - STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION - pp. da 233 a 241 ISBN: 9783642288937 | 9783642288944 ISSN: 1431-8814

M.C. Martini (2012) - *The refusal of offered jobs* (L. Fabbris - *Indicators of Higher Education Effectiveness*) (McGraw-Hill Milano ITA) - pp. da 49 a 60 ISBN: 9788838673306

8.Curriculum scientifico del P.I. (Max 3000 caratteri, spazi inclusi)

Professore associato confermato di Statistica Sociale presso il Dipartimento di Comunicazione ed Economia dell'Università di Modena e Reggio Emilia. Da aprile 2013 alla fine di ottobre 2015 ha fatto parte del Presidio per la Qualità dell'Ateneo di Modena e Reggio Emilia; questo periodo ha coinciso con il processo di

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

accreditamento dell'Ateneo. Dal 2015 è iscritta all'albo degli esperti della valutazione dell'Anvur (profilo esperti disciplinari).
Dal 2004 al 2005 è stata ricercatore di Statistica Sociale presso la Facoltà di Scienze Statistiche dell'Università di Padova. Ha conseguito nel 2002 il titolo di Dottore di Ricerca in Statistica Applicata alle Scienze Economiche e Sociali presso l'Università degli Studi di Padova, e nel 1998 la laurea in Scienze Statistiche e Demografiche.

L'attività scientifica è sviluppata principalmente nel settore della ricerca sociale, con particolare attenzione ai seguenti aspetti:
a. Analisi multivariata di dati economici e sociali
b. Valutazione della didattica universitaria e transizione scuola-lavoro
c. Analisi del rischio di povertà e disagio familiare
d. Analisi dei profili professionali tramite competenze e attività
e. Metodologie di indagine per la rilevazione di dati statistici

Ha partecipato ai gruppi di lavoro:
- "Servizi agli studenti e strumenti volti a ridurre abbandoni e ritardi all'università" (2010) per conto del CNVSU;
- "Analisi di sensibilità e sostituibilità di indicatori rilevanti per la valutazione del sistema universitario" (2005) per conto del CNVSU;
- "Sperimentazione di sistemi computer assisted per la rilevazione della valutazione della didattica universitaria da parte degli studenti e l'inserimento lavorativo, e professionale dei laureati e dei diplomati" (2000), per conto dell'Osservatorio per la valutazione del sistema universitario del MIUR.

Membro del collegio docenti della Scuola di dottorato in Medicina dello sviluppo e Scienze della programmazione sanitaria, indirizzo in Scienze della programmazione (Università di Padova), a partire dal XXIV ciclo.

Associate editor della rivista "Statistica Applicata – Italian Journal of Applied Statistics" dal 2011.

Revisore scientifico per la rivista "The European Journal of Development Research", Springer (dal 2013), per la rivista "Statistica Applicata – Italian Journal of Applied Statistics" (dal 2011), per la collana "Studies in Theoretical and Applied Statistics", Springer (dal 2011).

9.Abstract del progetto di ricerca (max 2000 caratteri, spazi inclusi)

L'analisi di dati testuali consiste in una serie di metodologie per estrarre informazioni da una raccolta di dati testuali; le analisi possono essere finalizzate a descrivere il lessico utilizzato nei testi, ad esaminare il contenuto dei testi ricostruito a partire dal lessico o da un processo di imputazione semantica, a combinare informazioni di carattere linguistico e interventi sul testo con analisi di tipo testuale, o ad estrarre da testi di grandi dimensioni informazioni rilevanti rispetto a specifiche interrogazioni. I testi analizzati possono essere i più disparati: titoli o articoli giornalistici, risposte a domande aperte, interviste strutturate, dichiarazioni e discorsi politici, testi pubblicitari, messaggi di reclamo o segnalazione di problemi, ma anche interventi sui social network (post di Facebook, tweet su Twitter), su forum di discussione, nelle chat.
Il progetto mira a costituire, presso il Dipartimento di Comunicazione ed Economia, una stazione mobile per l'analisi di dati testuali, attraverso l'acquisizione di un software specifico, di una postazione mobile su cui installare tale software, e di una biblioteca minima di testi scientifici e manuali relativi alla metodologia. La

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

postazione sarebbe poi messa a disposizione degli afferenti al Dipartimento per tutte le attività di ricerca che possono trarre beneficio dall'analisi di dati testuali. In questo senso la elevata multidisciplinarietà del Dipartimento costituisce un terreno fertile per l'introduzione di questo tipo di approccio, che unisce discipline presenti nel Dipartimento (la statistica, l'informatica, la linguistica) ed è di grande utilizzo in altre (il marketing, la psicologia, la sociologia, la semiotica, etc...).

10.Stato dell'arte (max 3000 caratteri, spazi inclusi)

L'analisi quantitativa dei dati testuali è costituita da un insieme di tecniche automatiche o semi-automatiche, supportate da software specifici, per la descrizione e l'analisi di dati testuali.
Dai primi approcci verso un'analisi quantitativa in ambito linguistico, risalenti agli anni Cinquanta e Sessanta con gli studi di Guiraud (1954, 1960) e Herdan (1964), si è passati attraverso il fondamentale sviluppo dell'analisi dei dati nel corso degli anni Settanta e Ottanta (Benzécri, 1977, 1982), che ha introdotto il concetto di statistica testuale basata sull'analisi di forme grafiche e e di segmenti ripetuti (Lebart e Salem, 1994; Lebart et al., 1998). Allo stesso tempo, sono stati sviluppati indici e misurazioni di statistica linguistica e statistica lessicale con le proposte di Muller (1977), Tournier (1980, 1985a, 1985b), Lafon (1980, 1981).
Negli anni Novanta, in Italia, si è assistito al crescente interesse per le misure di frequenza d'uso delle parole e allo sviluppo di dizionari elettronici e lessici di frequenza (De Mauro, 1998). Dagli studi stilometrici sull'opera di un autore e dalle analisi di testi letterari, l'interesse delle applicazioni si è spostato verso lo studio di corpora provenienti dalle fonti più diverse: raccolte di testi corti (titoli, testi pubblicitari), indagini sul campo in ambito psicologico e sociologico (domande aperte, interviste strutturate), analisi del discorso politico.
Gli sviluppi più recenti si incrociano con la diffusione dei cosiddetti "*big data*" (insiemi di dati così grandi in termini di volume, velocità di generazione e varietà da richiedere strumenti specifici per il loro trattamento): la diffusione del web 2.0 ha infatti generato una enorme quantità di testi in linguaggio naturale che circolano ogni istante su internet. L'analisi dei *big data* attraverso metodologie per dati testuali ha condotto allo sviluppo di tecniche di *opinion mining*, o *sentiment analysis* (vedi, ad esempio, Liu, 2012; Ceron et al., 2014), una metodologia per estrarre informazione dalle opinioni espresse dagli utenti sui social network.

Bibliografia essenziale:
Benzécri J.P. (1977) Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Données*, II, 4, pp.369-406.
Benzécri J.P. (1982) *Histoire et préhistoire de l'analyse des données*, Dunod, Paris.
Ceron A., Curini L., Iacus S.M. (2014) *Social media e sentiment analysis – L'evoluzione dei fenomeni sociali attraverso la Rete*, Springer-Verlag Italia, Milano.
De Mauro T. (1998) *Linguistica elementare*, Laterza, Roma-Bari.
Guiraud P. (1954) *Les caractères statistiques du vocabulaire*, PUF, Paris.
Guiraud P. (1960) *Problèmes et méthodes de la statistique linguistique*, PUF, Paris.
Herdan G. (1964) *Quantitative Linguistics*, Butterworths, London.
Lafon P. (1980) Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, I, pp. 127-165.
Lafon P. (1981) Analyse lexicométrique et recherche des cooccurrences, *Mots*, 2, pp.95-148.
Lebart L., Salem A. (1994) *Statistique textuelle*, Dunod, Paris.
Lebart L., Salem A., Berry L. (1998) *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.
Liu B. (2012) *Sentiment Analysis and Opinion Mining*, Graeme Hirst, Toronto
Muller Ch.(1977) *Principes et méthodes de statistique lexicale*, Hachette, Paris.
Tournier M. (1980) D'ou viennent les fréquences de vocabulaire?, *Mots*, I, pp.189-212.
Tournier M. (1985a) *Sur quoi pouvons-nous compter?* Hommage à Hélène Nais, Verbum.

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Tournier M. (1985b) Texte propagantiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation, *Mots*, II, pp.155-187

11.Ipotesi, obiettivi, metodologia e risultati attesi (max 8000 caratteri, spazi inclusi)

L'ipotesi alla base di questo progetto di creazione di una risorsa condivisa per l'analisi di dati testuali è che l'intrinseca multidisciplinarietà di questo approccio e delle sue applicazioni possa contribuire ad una diffusa creazione di valore all'interno di un Dipartimento a sua volta spiccatamente multidisciplinare. La centralità, in questo tipo di approccio, dell'analisi delle scelte lessico-testuali, appare particolarmente indicata all'interno di un Dipartimento che conta fra i suoi temi fondanti quello della comunicazione. Nonostante questo, e nonostante la vastissima estensione delle possibili applicazioni di analisi testuale e di *sentiment analysis* nell'ambito delle discipline presenti nel Dipartimento, questo approccio di analisi è ancora relativamente inesplorato tra i gruppi di ricerca e tra i ricercatori presenti, seppur con alcune eccezioni.
Operativamente, si intende partire dalla acquisizione e dalla predisposizione di una postazione mobile, su cui verrà installato un software specializzato per l'analisi di dati testuali. Il software da acquisire potrebbe essere TaLTaC^2, un software che consente l'analisi automatica di dati testuali secondo una duplice logica di analisi dei testi e di *text mining*. Tale software consente dunque un approccio quantitativo all'analisi di testi, sia per quanto riguarda la descrizione del linguaggio utilizzato, sia relativamente all'individuazione dei contenuti trattati nei testi.
Una volta "normalizzati" i testi da analizzare (standardizzando la grafia di parole e numeri e riconoscendo nomi, toponimi, e sigle, ma anche le principali locuzioni ed espressioni poliremiche), è possibile analizzarne il vocabolario calcolando diversi indici di ricchezza lessicale e frequenza delle forme grafiche. Successivamente, il riconoscimento delle forme grammaticali e la "lemmatizzazione" consentono di predisporre i testi per le successive analisi lessicali. L'analisi della sovra- o sotto-utilizzazione delle forme grafiche rispetto ad un vocabolario di riferimento consente di individuare forme specifiche del testo e parole-chiave; tra i possibili dizionari di frequenza a disposizione vi sono un dizionario dell'italiano standard (con un mix di forme dell'italiano scritto e parlato), un dizionario del lessico comune usato dalla stampa, e un dizionario del linguaggio economico-finanziario. Le risorse semantiche disponibili nel software sono, fra le altre, un dizionario di 6mila forme flesse di aggettivi categorizzati come positivi o negativi, utili per valutare la tonalità positiva o negativa di un testo (il cosiddetto *sentiment*), ma anche dizionari specifici (es. sull'enogastronomia, sulle locuzioni di luogo, sui crononimi). I risultati ottenuti in TaLTaC^2 possono essere utilizzati direttamente da altri software linguistici (es. Tree Tagger, Nooj-Intex, Lexical Studio) o statistici (es. SPAD, SPSS, SAS).
Accanto alle risorse informatiche, si intende acquisire e mettere a disposizione anche una piccola biblioteca specialistica contenente testi e manuali sull'analisi quantitativa dei dati testuali, sia da un punto di vista metodologico, sia relativamente alle numerose esperienze applicative nelle diverse discipline. Una prima parte dei testi verrà acquisita all'inizio del progetto, mentre l'acquisizione degli altri testi seguirà le esigenze che emergeranno nelle fasi successive.
Uno degli obiettivi del progetto è la disseminazione di conoscenze, anche metodologiche, fra i ricercatori delle diverse discipline; a tal fine si prevedono momenti di condivisione dell'impostazione, della conduzione e dei risultati delle analisi. Per garantire uno scambio fruttuoso verranno organizzati alcuni seminari interni al dipartimento, nel corso dei quali i primi utilizzatori delle nuove risorse di analisi di dati testuali potranno illustrare l'attività di ricerca in corso, al fine di introdurre i colleghi a questo approccio di analisi, mostrandone le potenzialità e le possibili applicazioni.

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

12.Eventuali potenzialità applicative e impatto scientifico e/o tecnologico e/o socio-economico (max 3000 caratteri spazi inclusi)

Il progetto mira a generare un impatto su due piani distinti: il primo legato all'ampliamento delle possibilità di analisi dei dati offerto dal Dipartimento, il secondo all'impatto scientifico nelle diverse discipline delle ricerche compiute da coloro che sceglieranno di avvalersi delle nuove risorse acquisite.
La disponibilità, senza costi per i singoli, di una risorsa condivisa e di una specifica biblioteca dedicata potrà infatti attrarre gli studiosi delle diverse discipline verso questo approccio di analisi ancora poco adottato nella realtà della ricerca dipartimentale; al contempo, la condivisione delle risorse e l'applicazione di una metodologia intrinsecamente multidisciplinare potranno favorire lo sviluppo di ricerche multidisciplinari all'interno del Dipartimento.
Parallelamente, sul piano dei contenuti disciplinari il progetto offre la possibilità di risultati interessanti in molteplici direzioni: certamente negli ambiti della statistica applicata, dell'informatica e della linguistica, che costituiscono il fondamento delle tecniche di analisi automatica e semiautomatica dei dati testuali, e dell'approccio lessico-testuale all'analisi dei testi in linguaggio naturale.
Le potenzialità, tuttavia, non sono limitate al possibile impatto nelle discipline fondative di queste tecniche, ma si possono estendere a tutte quelle tematiche che possono trarre vantaggio dalla possibilità di estrarre informazione da grandi quantità di testi scritti o da qualsiasi testo espresso in linguaggio naturale. Ne sono esempi, fra gli altri, le discipline che si avvalgono di *case studies* basati, fra le altre cose, su interviste a testimoni privilegiati oppure a *stakeholders*, ma anche tutte le discipline sociali che prevedono il ricorso, accanto ai questionari strutturati con domande chiuse, anche a interviste strutturate e questionari con domande aperte. L'impatto risulta ancora più forte quando si pensa alla possibilità di analizzare dati non esplicitamente sollecitati attraverso questionari o interviste, ma forniti spontaneamente dagli utenti dei social network; queste applicazioni sono di grande interesse negli studi elettorali, per i quali negli ultimi anni le ricerche di tipo tradizionale hanno mostrato serie difficoltà nel cogliere il reale orientamento degli elettori, ma anche le analisi di marketing, per le quali l'esame tramite *sentiment analysis* delle conversazioni online su un determinato brand, prodotto o servizio possono aiutare a determinarne posizionamento, percezione e valutazione.
Un risultato positivo della sperimentazione prevista da questo progetto potrà eventualmente suggerire di proseguire l'utilizzo delle risorse con investimenti da parte di singoli gruppi di ricerca o del Dipartimento.

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

13.Costo complessivo del progetto articolato per voci di costo:

	Costo	Descrizione Max. 2.000 caratteri spazi inclusi
Eventuale cofinanziamento (certificato dal dipartimento)		
Costo dei contratti del personale da reclutare		
Attrezzature, strumentazioni e prodotti software	2250	Acquisto di un software per l'analisi di dati testuali e di un computer portatile
Servizi di consulenza e simili		
Altri costi di esercizio (missioni, partecipazioni a convegni, attività di disseminazione dei risultati, pubblicazioni, organizzazione convegni, seminari, materiale di consumo, ecc)	1500	Acquisto di testi scientifici e manuali relativi all'analisi di dati testuali; materiale di consumo; disseminazione dei risultati
Totale	3750	

Data, 11/11/16

Firma del Responsabile scientifico
M. Chiara Meluzzi