

**ALMA MATER STUDIORUM – UNIVERSITA’
DIBOLOGNA**

DIPARTIMENTO DI SCIENZE STATISTICHE

“PAOLO FORTUNATI”

Corso di Laurea in Scienze Statistiche

Curriculum Economia ed Impresa

Analisi testuale e sentimentale sul mondo delle autovetture

Presentata da:
Tommaso Possenti
Matricola: 001002373

Relatore:
Prof. Luca Trapin

APPELLO III

ANNO ACCADEMICO 2022/2023

Indice

1. INTRODUZIONE.....	2
2. DESCRIZIONE ED ANALISI PRELIMINARE DEI DATI	5
2.1 DATASET	5
2.2 PROCEDIMENTO INIZIALE.....	6
2.2.1 <i>Prima interazione con i dati</i>	6
2.2.2 <i>Pulizia e riorganizzazione dei dati</i>	6
3. TEXT MINING E SENTIMENT ANALYSIS	7
4. ANALISI EMPIRICA.....	12
4.1 È POSSIBILE TRAMITE UN'ANALISI TESTUALE ESEGUITA SU RECENSIONI INDIVIDUARE CARATTERISTICHE POSITIVE E NEGATIVE ASSOCIATE A MODELLI DI AUTOMOBILI?	12
4.2 È POSSIBILE PARTENDO DA COMMENTI VALUTARE QUANTO LE PERCEZIONI DEL CONSUMATORE SULLE CARATTERISTICHE DI UN'AUTO IMPATTINO SUL VOTO TOTALE?	21
4.3 RICORRENDO ALL'UTILIZZO DI FUNZIONI DI TEXT MINING E SENTIMENT ANALYSIS È POSSIBILE COMPRENDERE LE CAUSE DELL'ANDAMENTO DEI PUNTEGGI MEDI DI UN MODELLO NEGLI ANNI?	25
4.3.1 <i>Rilevamento punteggi medi</i>	25
4.3.2 <i>Rilevamento ed analisi delle frequenze più utilizzate e degli spot pubblicitari</i>	28
4.3.3 <i>Conclusione</i>	32
5. CONCLUSIONI	33
6. BIBLIOGRAFIA	34

1. Introduzione

L’obiettivo di questo elaborato è rappresentato dalla curiosità di analizzare e rappresentare il mercato delle auto attraverso gli occhi dei consumatori, grazie alle recensioni rilasciate sul sito Edmunds.

All’interno di questo sito web sono contenute informazioni e recensioni su numerosi modelli di automobili in commercio negli United States.

L’analisi e la rappresentazione del mercato delle automobili si concretizzano all’interno di tre domande principali che racchiudono alcuni degli aspetti più rilevanti dell’Analisi di mercato.

Nello specifico si vuole verificare se è possibile, tramite text e sentiment analysis:

1. Identificare le caratteristiche positive e negative principali che il consumatore associa a prodotti di marca differente.
2. Quantificare l’impatto che le percezioni di determinati attributi hanno sulla valutazione complessiva di un prodotto.
3. Risalire alle caratteristiche che hanno determinato l’andamento delle valutazioni del prodotto negli anni.

Con questa tesi si vogliono esporre i risultati ottenuti dall’applicazione di tecniche di text mining e sentiment analysis su di un dataset formato da recensioni e valutazioni appartenenti al mondo del mercato automobilistico.

L’analisi di tale mercato risulta essere particolarmente interessante poiché le automobili ed il loro acquisto rappresentano una parte centrale della sfera socioeconomica dell’uomo moderno.

L’automobile è un bene la cui scelta rappresenta il risultato di una ricerca ragionata ed approfondita da parte del consumatore, essendo un acquisto poco frequente e dal costo elevato.

L’acquisto di un modello è dunque successivo ad una fase di valutazione logica ed oggettiva da parte del consumatore; pertanto, saranno le caratteristiche ed i benefici associati all’auto a definirne il successo in termini di vendite.

Inoltre, tale mercato, risulta essere ideale per l’analisi che si andrà a sviluppare date le sue dimensioni a livello di: consumatori, di modelli di automobili disponibili e di livelli

di attributi di cui ogni modello possiede combinazioni differenti.

Il dataset considerato contiene recensioni e valutazioni espresse dai consumatori su numerosi marchi e modelli disponibili in commercio tra l'anno 2000 ed il 2020.

Si è deciso di utilizzare un dataset formato prevalentemente da commenti poiché questi rappresentano i riscontri principali che un'azienda riceve dai propri consumatori e sono portatori di ampie quantità di informazioni.

Tuttavia, i commenti, essendo sviluppati direttamente dal consumatore, contengono informazioni che risultano essere latenti e pertanto per estrarle è necessario lavorare i testi in modo da uniformarli e renderli analizzabili.

Sfruttando gli strumenti del Text mining ed affiancandoli a funzioni di Sentiment analysis è stato possibile rilevare le percezioni dei consumatori nei confronti delle automobili e dei loro attributi principali.

L'elaborato prevede tre domande cardine che rappresentano alcuni degli aspetti più rilevanti all'interno di un'analisi di mercato.

Nella prima domanda: **“È possibile tramite un’analisi testuale eseguita su recensioni individuare caratteristiche positive e negative associate a modelli di automobili?”** ci si è chiesti se fosse possibile, utilizzando come input i commenti, ottenere informazioni utili allo sviluppo di una prima panoramica del mercato delle auto.

Più nello specifico, ci si è chiesti se fosse possibile identificare i punti di forza e di debolezza dei modelli agli occhi del cliente, oltre ad eventuali difetti e problemi riscontrati dai consumatori; tutte informazioni fondamentali in termini di marketing e miglioramento del prodotto.

Per fare ciò si sono valutate le frequenze di singole parole e coppie di espressioni all'interno dei commenti, con l'obiettivo di ritrovare le espressioni più utilizzate nella descrizione di ogni modello.

Con questa domanda si vogliono rilevare le percezioni dei consumatori associate alle caratteristiche dei vari modelli, si sono ricercate le caratteristiche di un'automobile maggiormente associate a determinati termini, sia positivi che negativi, come: “good”, “bad”, “problem”.

Si è, ad esempio, considerata l'espressione “problem” e l'algoritmo ha individuato le caratteristiche maggiormente associate a questo termine all'interno dei commenti.

Questo procedimento ha permesso di ottenere una panoramica delle percezioni che il

consumatore ha sviluppato verso un modello ed i suoi attributi.

Nella seconda domanda: “**È possibile partendo da commenti valutare quanto le percezioni del consumatore sulle caratteristiche di un auto impattano sul voto totale?**” si è valutato quanto la percezione positiva o negativa dei consumatori rispetto ad alcune caratteristiche di un modello, abbia impattato sul punteggio finale rilasciato nelle recensioni.

In quest’analisi si sono sfruttate le informazioni raccolte nell’analisi delle frequenze per definire quattro caratteristiche principali “look”, “drive”, “Comfort” e “transmission”. Partendo proprio da queste quattro caratteristiche si è impostata un’analisi di regressione multipla, in cui le caratteristiche considerate rappresentano le variabili indipendenti. Mentre, la variabile dipendente è rappresentata dal punteggio finale rilasciato dai consumatori all’interno di ogni recensione.

Con questa domanda si vuole valutare l’impatto che la percezione del consumatore, rispetto a quattro caratteristiche dell’automobile, ha sulla valutazione finale, con l’obiettivo di verificare se, partendo da commenti, sia possa sviluppare un’analisi in grado di individuare le caratteristiche più importanti per il consumatore e di quantificarne la rilevanza.

Nella terza domanda: “**Ricorrendo all’utilizzo di funzioni di text mining e sentiment analysis è possibile comprendere le cause dell’andamento dei punteggi medi di un modello negli anni?**”

Si è rappresentata graficamente la serie storica dei punteggi medi annuali rilasciati dai consumatori per il modello Focus del marchio Ford.

Successivamente si è valutata anche la serie storica dei punteggi sentimentali medi annuali rilevati nei commenti per il medesimo modello.

Si sono poi rilevate le espressioni positive e negative più utilizzate in ogni anno, andando così a ritrovare le cause dell’andamento delle serie storiche negli anni.

Infine, sono stati ritrovati gli spot pubblicitari degli anni considerati per poter valutare l’attività di marketing utilizzata dal marchio Ford rispetto all’andamento delle serie storiche.

Con questa domanda si vuole valutare l’evoluzione del modello Ford Focus negli anni, attraverso gli occhi del consumatore e studiare le strategie di marketing utilizzate.

La tesi è strutturata nella maniera seguente:

Nel capitolo due si approfondiranno le tecniche utilizzate per l'avvio della ricerca in termini di esplorazione e preparazione dei dati per le successive analisi.

Nel capitolo tre si andranno a esporre i concetti di text mining e sentiment analisi, approfondendo gli algoritmi e le funzioni utilizzati nella parte empirica dell'analisi.

Nel capitolo quattro si ritroveranno le domande sopra descritte assieme ad output del software r riportati sotto forma di grafici o tabelle utilizzati durante l'estrazione delle informazioni.

L'elaborato termina con il capitolo cinque e sei contenenti rispettivamente conclusioni finali e bibliografia/sitografia.

2. Descrizione ed analisi preliminare dei dati

2.1 Dataset

Il dataset di partenza è stato estrapolato dal sito web KAGGLE, una piattaforma di competizione e collaborazione nell'ambito della “data science” e del “machine learning”.

Il dataset individuato era inizialmente denominato: “Edmund car review”, come suggerisce il nome stesso i dati al suo interno provengono dall'archivio digitale e sito web: “Edmund”, che costituisce un'importante risorsa digitale per l'industria automobilistica, sia sotto forma di archivio di caratteristiche fisiche delle auto entrate in circolazione che di recensioni e commenti particolarmente ampi sulle automobili stesse.

Il dataset è stato creato tramite un'operazione di web-scraping articolata attorno a 46 marchi di automobili ad esempio, BMW, Alfa-Romeo, Honda, Fiat, per un totale di 897 modelli messi in commercio nell'arco temporale che va dal 2000 al 2019.

La forma del dataset è costituita da otto colonne: Compagnia, modello, anno, recensore, data, titolo, punteggio, recensione; da 299'045 righe, ciascuna delle quali contiene le informazioni associate ad un singolo commento.

Per quanto riguarda la colonna punteggio essa contiene una valutazione su scala da zero a cinque in cui i consumatori riportano un voto complessivo riferito al modello considerato.

Mentre la colonna recensione contiene i commenti rilasciate da ogni utente o

consumatore.

2.2 Procedimento iniziale

2.2.1 Prima interazione con i dati

Individuato il dataset di partenza e dopo averlo importato sul software “RStudio” si è iniziata una prima fase di familiarizzazione con i dati e di esplorazione all’interno del dataset stesso al fine di valutare la complessiva correttezza formale del contenuto e presenza di commenti o altri campi vuoti (“NA”).

Il dataset si è presentato complessivamente privo di forme errate o assenti.

Ha poi fatto seguito una fase di valutazione a campione della correttezza e qualità dei commenti, al fine di ritrovare eventuali commenti articolati in lingue differenti da quella inglese e poter stabilire se i commenti vi fossero sufficienti informazioni per portare avanti l’intera analisi.

2.2.2 Pulizia e riorganizzazione dei dati

Verificata la complessiva validità ed utilità dei dati si è passati alla stesura della prima parte del codice R denominata “Scrematura&Pulizia”, che per l’appunto vede presentarsi due fasi differenti.

Nella prima parte del codice, il dataset è stato ridotto rispetto alle variabili: “Anno”, “Modello” e “Marchio”; nella prima parte dell’analisi sono di fatto stati considerati quattro modelli di auto, Ford Focus, Hyundai Tucson, Toyota Prius, Toyota Highlander, all’interno dell’anno 2016, tutto ciò al fine di velocizzare ed ottimizzare il tempo di analisi necessario al sistema per elaborare i dati nelle fasi successive.

Nella seconda parte del codice si sono introdotte funzioni ed algoritmi appartenenti al mondo dell’analisi testuale, ovvero del Text Mining.

Nello specifico, si è introdotta una fase di pulizia del testo, in cui sono stati eliminati dai commenti contenuti non utili all’analisi, come: numeri, punteggiatura, caratteri maiuscolo e simboli.

Successivamente sono state eliminate espressioni poco utili a causa della bassa informazione trasportata e dell’utilizzo troppo frequente, come: “macchina”, “comprare” e “avere”.

Dopo aver ripulito i testi, si è passato a trasformarli in forma strutturata e

standardizzata, operativamente si sono ricondotte tutte le espressioni alla loro forma base in modo da ottenere testi omogenei.

Espressioni come verbi, aggettivi e nomi sono stati privati di coniugazioni e desinenze e sono stati ricondotti alla loro radice.

3. Text mining e Sentiment Analysis

Il text mining rappresenta un procedimento di estrapolazione di informazioni latenti e di pattern nascosti contenuti all'interno di righe di testo, tramite un'analisi automatizzata.

Il text mining è inoltre alla base di servizi fondamentali che le aziende offrono ai propri consumatori, si consideri ad esempio il servizio clienti diventato completamente automatizzato grazie ai chat bot.

Tali software o bot rilevano e comprendono i problemi espressi dai consumatori tramite l'analisi del testo.

Nello specifico il text mining è un insieme di funzioni in grado di ripulire il testo da termini inutili o irrilevanti per l'analisi e di standardizzare le espressioni in modo da consentire lo sviluppo di ulteriori analisi come quella sentimentale.

Operativamente, partendo da commenti si procede a pulirli da eventuali errori o elementi indesiderati, si esegue una modifica di alcune espressioni che in certi casi vengono rimosse o vengono identificate come forme composte, come ad esempio: “road noise” oppure “mile per gallon”.

Il text mining è associabile al processo inverso della costruzione di un muro di mattoni.

Il testo di partenza rappresenta il muro completo e finito, attraverso algoritmi e funzioni si punta ad analizzare i singoli mattoni del muro, ovvero le singole espressioni o parole del testo.

Si estraggono e si valutano i singoli mattoni in modo da comprendere ogni loro caratteristica e quindi di poter comprendere a pieno le caratteristiche stesse del muro di mattoni e potere estrarre informazioni utili all'analisi.

All'interno del mondo testuale, sul quale in text mining si basa, ritroviamo tre livelli differenti:

- a) Livello macro: il corpus, che può essere immaginato come un libro o nel nostro caso il quasi milione di recensioni disponibili
- b) Livello intermedio: i testi, che possiamo pensarli come i capitoli e quindi, nel caso attuale, la singola recensione
- c) Livello micro: le parole

L’analisi e lo sviluppo degli algoritmi, parte dal livello intermedio nel quale ogni testo viene pulito attraverso varie fasi che hanno il compito di eliminare eventuali elementi non desiderati o considerati irrilevanti per l’analisi, quali: hashtag, numeri, menzioni.

In seguito, l’algoritmo procede ad eliminare dal testo eventuali espressioni prive di informazioni utili per l’analisi data la loro natura logico-grammaticale, espressioni denominate “stopword”.

Le espressioni che ricadono all’interno di questa categoria vengono identificate automaticamente dall’algoritmo tramite l’utilizzo di dizionari appositi per la lingua del testo in analisi.

Nel dataset considerato in questo elaborato si sono ritrovate stopword come: “I”, “the”, “if”, “then”.

A queste espressioni rilevate tramite dizionari l’utente può aggiungere delle parole che considera irrilevanti o troppo frequenti e pertanto prive di utilità, tra le parole così individuate vi sono: “car”, “vehicle”, “honda” e “toyota”.

A questa fase di pulizia fa seguito una fase di riorganizzazione del testo, che ha come scopo l’individuazione e la correzione di forme composte, ovvero combinazioni di più parole che formano un’espressione unica e a sé stante.

Ad esempio, l’espressione “ruota di scorta” pur essendo formata da più parole può essere considerata come una singola espressione e pertanto l’algoritmo la modificherà in “ruota_di_scorta”.

Quest’operazione ha l’obiettivo di preservare l’informazione trasportata dalle varie espressioni così da semplificare e rendere più affidabile l’analisi.

Anche in questa fase si fa ricorso sia a dizionari digitali, sia ad espressioni introdotte direttamente dall’utente, come ad esempio: “steering wheel” che viene riscritta come “steering_wheel”.

La trasformazione delle forme composte ci permette di facilitare l’interpretazione del testo e la ricerca del senso logico della frase.

Alla pulizia e riorganizzazione del testo fa poi seguito una fase detta di lemmatizzazione. In questa fase ogni parola del testo viene ricondotta al suo lemma, dove con lemma si intende la combinazione di vocabolo e categoria grammaticale.

Questa operazione ci permette di normalizzare coniugazioni differenti di verbi o la declinazione di nomi, aggettivi, pronomi, ad esempio: di fronte alle parole “cars”, “drove” l’output in seguito alla lemmatizzazione sarà “car, NOUN”, “drive, VERB”.

Terminata la fase di Text mining, si procede all’introduzione delle funzioni di analisi sentimentale.

Tramite la Sentiment Analysis si è reso possibile identificare l’emozione contenuta all’interno di un testo, ovvero comprende il tono con cui un consumatore ha rilasciato una determinata recensione.

L’obiettivo primario di tale analisi consiste nell’identificare il testo e le espressioni in esso contenute come positive, negative o neutre.

Per fare ciò si sfruttano dei dizionari digitali tramite i quali il software è in grado di identificare le espressioni associabili a stati d’animo come, ad esempio ira, gioia, sorpresa, frustrazione, all’interno dei commenti;

Ogni emozione rilevata nei commenti ha un proprio punteggio sentimentale che può essere maggiore o minore di zero e quindi una generale accezione positiva o negativa.

La somma dei punteggi sentimentali delle espressioni che formano un commento definisce il punteggio sentimentale complessivo della recensione.

Si può quindi sviluppare uno studio delle emozioni e dei sentimenti riportati all’interno dei commenti.

Da un punto di vista procedurale, si sono considerati due dataset di partenza risultanti: uno dalla pulizia del testo, l’altro dalla lemmatizzazione dei testi puliti.

Il primo di questi dataset contiene su ogni riga un commento pulito da simboli, punteggiatura, maiuscole, forme composte e stop-word.

Mentre nel secondo dataset si ritrovano su ogni riga le singole espressioni che costituiscono ciascun commento, ciascuna etichettata in base al commento di appartenenza.

In questo secondo dataset ad ogni espressione si affianca la natura morfologia dell’espressione, come ad esempio nome, verbo, pronome o aggettivo.

Sfruttando il primo dataset si è eseguita un’analisi che ha portato alla definizione del punteggio sentimentale associato ad ogni commento.

Tale operazione si è svolta facendo ricorso alla funzione R “get_nrc_sentiment ” del pacchetto “syuhet”.

L’algoritmo procede ad analizzare il testo e fornisce come output un dataset in cui ogni riga corrisponde ad un testo e le colonne sono così denominate: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive".

L’algoritmo riporta sotto ogni colonna il numero di espressioni, rilevate nei commenti, che corrispondono all’emozione titolo della colonna.

In pratica, questa funzione riporta quante volte, in un commento, è stata utilizzata ciascuna emozione.

In base alla frequenza di ogni emozione l’algoritmo ricava il numero di emozioni positive e negative riportate in ogni commento, questo all’interno delle ultime due colonne del dataset.

Andando a sommare l’opposto del punteggio delle emozioni negative al punteggio delle emozioni positive, si ottiene la “valenza” dell’intero testo, ovvero il punteggio sentimentale dell’intero commento.

Così facendo si può comprendere il tono complessivo con cui è stato sviluppato ogni commento.

Si è poi valutata la relazione tra la media dei punteggi rilasciati direttamente nelle valutazioni (asse delle ascisse) e i punteggi sentimentali medi (asse delle ordinate) facendo ricorso ad una rappresentazione grafica di tipo scatter-plot sui dati relativi a sedici modelli nell’anno 2014.

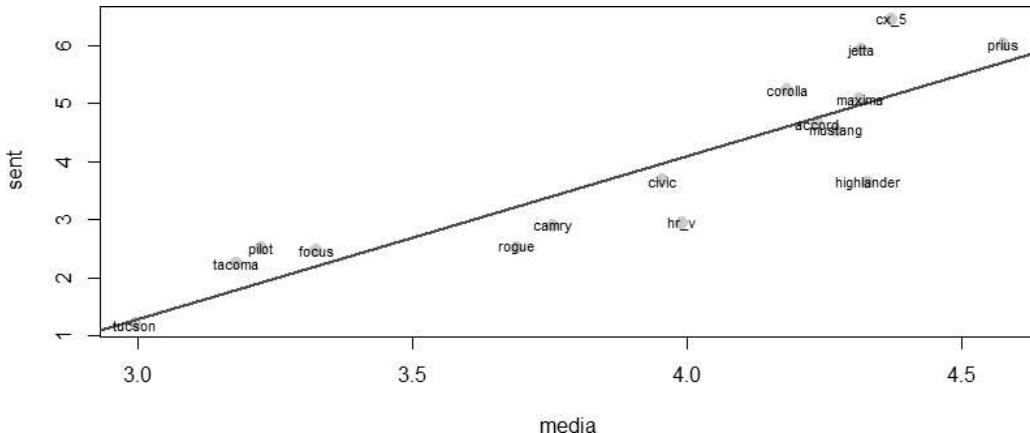


Figura 1. Scatter-plot di 16 modelli di automobili. Sull’asse x si ritrovano i punteggi medi calcolati per ogni modello nell’anno 2014. Sull’asse y sono riportati i punteggi sentimentali medi rilevati nei commenti di ogni modello nell’anno 2014.

In **Figura 1** si noti la retta in rosso, che rappresenta l’ideale rapporto tra punteggi rilasciati medi e punteggi sentimentali medi.

Si nota subito come la nube dei punti formata dai modelli sia prevalentemente compatta e aderente alla retta, confermando il logico collegamento tra punteggi rilasciati e punteggi sentimentali calcolati.

Tuttavia, si vedono alcuni modelli che si allontanano più di altri dalla retta, come la Mazda cx_5, che presenta un punteggio medio di circa 4.4 che dovrebbe corrispondere ad un punteggio sentimentale pari a circa 5, ma che nella realtà risulta essere superiore a 6.

Questo risultato evidenzia l’importanza dell’analisi sentimentale in termini di informazione aggiuntiva che si ha la possibilità di estrarre dai commenti.

Per quanto riguarda il secondo dataset, questo è stato sfruttato per eseguire l’analisi sentimentale su sottotesti che contenessero specifiche espressioni.

Infatti, sono stati individuate parti di testo che contenessero espressioni come “look”, “Comfort” e “transmission” e si andati a valutare le emozioni maggiormente riportante all’interno di questi sottotesti.

Attraverso questo procedimento si è potuto sviluppare una ricerca del punteggio sentimentale associato a specifiche caratteristiche di un’automobile e quindi si è portata avanti un’analisi che avesse come obiettivo quello di valutare l’importanza percepita di alcune caratteristiche dell’automobile.

4. Analisi empirica

4.1 È possibile tramite un’analisi testuale eseguita su recensioni individuare caratteristiche positive e negative associate a modelli di automobili?

Con questa domanda si punta a ritrovare le caratteristiche positive e negative percepite dai consumatori che caratterizzano maggiormente i vari modelli.

Nello specifico si vuole sviluppare una tabella che riassume le caratteristiche maggiormente citate all’interno dei commenti di quattro modelli di automobili differenti: Ford Focus, Hyundai Tucson, Toyota Prius e Toyota Highlander nell’anno 2014.

Questi quattro modelli sono stati individuati sulla base di due criteri: la numerosità in termini di commenti e la posizione all’interno del grafico in **Figura 1**; con l’obiettivo di considerare modelli differenti e variegati in termini di punteggi medi rilasciati e possibilmente di caratteristiche positive e negative associate.

Alla base dell’analisi di questo paragrafo vi è la valutazione delle frequenze delle espressioni contenute nei commenti.

Nel corso dell’analisi, si utilizzeranno metodi differenti per la rilevazione delle frequenze e per ciascuno di questi metodi verrà sviluppata una tabella dei termini dalla frequenza più elevata.

La tabella finale, risultante dal perfezionamento dell’analisi, rappresenterà una panoramica riassuntiva del mercato delle automobili dal punto di vista del consumatore.

In questa tabella verranno riportate: le espressioni più associate ad ogni modello, le caratteristiche positive e negative dall’utilizzo più frequente e le espressioni maggiormente affiancate a termini come “problem” o “issue”.

Lo scopo di questa tabella è rappresentare le percezioni dei modelli attraverso gli occhi e le esperienze del cliente.

Lo scopo di quest’analisi consiste nell’individuare i punti di forza e debolezza dei vari modelli, individuando possibili miglioramenti del prodotto che soddisfino i bisogni del cliente .

Il primo passaggio all’interno di quest’analisi è stato valutare il numero di commenti disponibili per ogni modello nell’anno 2014, che si è visto essere pari a circa duecento per ogni modello considerato.

Partendo dal dataset costituito da questi commenti si è proceduto ad utilizzare una

funzione “filter” per estrapolare i dati relativi ai modelli scelti nell’anno considerato. Prima di sviluppare l’analisi delle frequenze è stato necessario eseguire quattro operazioni di text mining mirate ad eliminare o perlomeno limitare altrettanti problemi principali.

Come primo passaggio i testi sono stati ricondotti al medesimo codice di codifica, in questo caso è stata utilizzata la codifica UTF-8.

Il motivo di questa operazione è quello di rendere i testi uniformi ed evitare che commenti con codifiche particolari non vengano propriamente visualizzati dal software. All’interno di questa fase si sono anche ripuliti i testi da elementi come numeri, simboli e maiuscole.

Come secondo passaggio si è dovuto standardizzare le espressioni all’interno dei testi, ogni espressione è stata ricondotta alla sua forma di base in modo da eliminare eventuali coniugazioni o desinenze che altrimenti avrebbero sfalsato la valutazione delle frequenze.

Tramite questo procedimento i testi vengono trasformati in modo da renderli tra loro confrontabili ed uniformi.

Ad esempio, le espressioni “drive”, “drove”, “driven” che sono tra loro differenti ma dal significato analogo, vengono ricondotte alla forma base del verbo e pertanto vengono tutte trasformate in “drive”.

Il terzo passaggio è stato eliminare dai testi le espressioni denominate “stop-word”.

Tali espressioni possono essere pronomi, articoli o espressioni dall’utilizzo frequente ma dalla bassa se non nulla informazione trasportata nell’ambito di un’analisi sentimentale o testuale.

Il compito di queste espressioni nella lingua parlata e scritta è quello di collegare i vari elementi di una frase, per renderne completo il significato.

Tali termini sono stati rimossi utilizzando un dizionario apposito di stop-word della lingua inglese, che è stato inserito all’interno dell’algoritmo utilizzato per la pulizia dei testi.

Anche queste espressioni possono sfalsare la rilevazione delle frequenze poiché possono ottenere facilmente frequenze molto elevate.

Importante notare come sotto la categoria di stop-word sono state inserite manualmente

delle espressioni proprie del mondo automobilistico come, ad esempio, i termini “car” o “buy”, espressioni che non trasportano alcuna informazione aggiuntiva per l’analisi.

Il quarto passaggio è stato individuare e contrassegnare le espressioni in forma composta, ovvero combinazioni di due, tre o anche quattro parole in grado di formare un’espressione dal senso completo e a sé stante.

Esempi di queste espressioni possono essere “ruota di scorta” o “sedile del passeggero”. Anche qui sono stati utilizzati dizionari e interazioni manuali per rilevare tali espressioni e segnarle come espressioni uniche e non come parole separate.

Questo per evitare la perdita delle informazioni trasportate all’interno di tali forme composte.

Eseguiti questi primi procedimenti si è passati alla rilevazione delle frequenze.

Nella valutazione delle frequenze delle espressioni si sono utilizzati metodi differenti per l’assegnazione dei pesi.

Inizialmente si è utilizzata la ponderazione Tf.

Attraverso questo primo metodo si è valutata la “term-frequency”, ovvero la frequenza con la quale si presentano determinate espressioni all’interno dei commenti, cioè il numero di volte che un termine compare in tutti i commenti.

Questo primo metodo di ponderazione, tuttavia, comportava due problemi principali.

Un’eccessiva importanza assegnata ad espressioni non categorizzabili come stop-word, ma che comunque presentano frequenze elevate, come nel caso del termine “get” che è utilizzato nella maggior parte dei testi ma non è portatore di alcuna informazione rilevante.

Il secondo problema è la perdita di informazioni dovuta alla mancata suddivisione dei commenti in postivi e negativi.

Questo problema si basa sull’impossibilità di comprendere il motivo per cui una certa espressione è citata molte volte nei commenti.

Infatti, senza una suddivisione tra commenti positivi e negativi, non è possibile comprendere il senso vero e proprio delle frequenze o, meglio, non si è grado di risalire

alla causa di un utilizzo elevato di un determinato termine.

Ad esempio, si è spesso presentata la situazione in cui vi fosse un elevato utilizzo dell'espressione "transmission" senza poter dire se la trasmissione fosse citata spesso grazie ad un'ottima qualità meccanica oppure a causa di problemi tecnici e difetti ad essa legati.

Nella risoluzione di questi problemi sono state individuati due metodi differenti.

1. Frequenza delle espressioni con ponderazione tf-idf

Per porre rimedio alle mancanze ed i difetti del metodo tf si è passati ad utilizzare un nuovo metodo di ponderazione, il metodo Tf-idf.

Dove Tf sta per "term frequency" mentre idf rappresenta l'acronimo di "inverse document frequency", ovvero l'inverso dei numeri di documenti in cui una certa parola compare.

Con questa nuova ponderazione si affida un punteggio più elevato ai termini che non si presentano in troppi commenti.

Questo metodo va a premiare le espressioni che hanno un valore maggiore in termini di informazione trasportata, nonostante la bassa frequenza.

Mentre va a limitare l'importanza di termini dalla frequenza elevata ma che non presentano un valore semantico rilevante, come nel caso dell'espressione "get".

In pratica, questo metodo va a ridurre la frequenza associata a termini che si ritrovano in un numero troppo elevato di commenti.

Al nuovo metodo di ponderazione tf-idf è stata poi associata una suddivisione dei commenti in positivi e negativi.

Sono stati considerati come positivi i commenti aventi punteggio sentimentale pari o superiore a zero e come negativi quelli aventi punteggio sentimentale inferiore a zero.

Così facendo è stato possibile sviluppare una ricerca delle frequenze che consentisse di assegnare una rilevanza ed un significato giusto e corretto ai risultati rilevati.

2. Co-occorrenza partendo da espressione predefinita

Questo metodo si allontana dalla valutazione delle frequenze delle singole espressioni e si concentra sulla valutazione delle co-occorrenze, ovvero sulla rilevazione delle frequenze di coppie di espressioni.

L’obiettivo di questo metodo è valutare direttamente le associazioni tra espressioni dall’accezione positiva o negativa e le caratteristiche di un’automobile.

Individuando le caratteristiche dell’auto più spesso riportate assieme a termini positivi si possono individuare gli attributi maggiormente apprezzati dai consumatori.

Nello specifico il codice che è stato sviluppato richiede, come matrice di partenza, la matrice costituita dalle co-occorrenze di tutte le espressioni all’interno dei commenti.

Dopo aver definito un termine di partenza, l’algoritmo rileva tutte le coppie di espressioni in cui è contenuto tale termine e le ordina per frequenza in senso decrescente.

Modello	1°	2°	3°	4°	5°	6°
Prius	Get	Age	Wheel	Look	Drive	Better
Focus	Transmission	Get	Problem	Issue	Start	Clutch
Highlander	Seat	Love	Build	Good	Purchase	Quality
Tucson	Bad	Old	Get	Seat	Drive	Better

Tabella 2, Tabella dei modelli Prius, Focus, Highlander e Tucson. La prima colonna riporta i modelli considerati e le sei colonne successive riportano le sei espressioni dalla frequenza più elevata per ogni modello rilevate con metodo di ponderazione tf ed ordinate in senso decrescente.

La **Tabella 2** le espressioni dalla frequenza più elevata nei commenti dei quattro modelli considerati.

In questa tabella sono riportate le sei espressioni dalla frequenza più elevata rilevate nei commenti di ogni modello tramite ponderazione tf e senza suddivisione tra commenti positivi e negativi.

Lo scopo di questa tabella è quello di mostrare i problemi associati al metodo tf inizialmente descritti.

Infatti, per ciascun modello si vedono utilizzate espressioni come “get” o aggettivi che non forniscono le informazioni che si ricercano in quest’analisi.

Modelli	1°	2°	3°	4°	5°	6°
Prius	Great	Love	Seat	Handle	Comfort	Quality
Focus	Problem	Transmission	Fix	sync	Good	Issue
Highlander	Seat	Love	Mpg	Price	Better	Quality
Tucson	Dislike	Bad	Problem	Seat	Transmission	Better

Tabella 3. Tabella dei modelli Prius, Focus, Highlander e Tucson. La prima colonna riporta i modelli considerati e le sei colonne successive riportato le sei espressioni dalla frequenza più elevata per ogni modello rilevate con metodo di ponderazione tf-idf ed ordinate in senso decrescente.

Nella **Tabella 3** si mostrano i risultati ottenuti della sostituzione al metodo tf del metodo tf-idf.

Confrontando la tabella appena ottenuta con la **Tabella 2** si può comprendere la differenza di risultati ottenuti con i due metodi.

Attraverso il metodo idf otteniamo risultati differenti ed apparentemente più puliti e utili di quelli ottenuti col metodo tf.

Si perdono quei termini caratterizzati da una frequenza molto elevata in ogni commento come, ad esempio, il termine “Get” che lascia i primi sei posti della tabella.

L’analisi con metodo idf è stata eseguita su nomi, verbi e aggettivi, in maniera congiunta; tuttavia, per approfondire l’analisi delle caratteristiche positive e negative associate ad ogni modello si è deciso di limitare l’analisi delle frequenze con metodo idf ai soli nomi. Questo per centrare l’attenzione della ricerca sugli attributi dei modelli, in modo da individuare esclusivamente le caratteristiche maggiormente commentate, andando poi a suddividere i commenti in positivi e negativi.

La suddivisione dei commenti tra negativi e positivi è stata impostata andando a utilizzare le funzioni della sentiment analysis.

Come criterio di discriminazione tra commenti positivi e negativi si è utilizzata la “valenza” dei commenti, ovvero il punteggio sentimentale complessivo del commento, risultante dalla somma di tutti i punteggi sentimentali associati ad ogni singola espressione del testo.

Per calcolare la valenza si è sviluppato un algoritmo che, partendo dal dataset dei testi lemmatizzati e da dizionari di emozioni in lingua inglese, restituisse una matrice avente per righe i singoli commenti e per colonne tutte le seguenti emozioni: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative",

"positive".

Successivamente, attraverso la funzione di R “get_nrc_sentiment” del pacchetto “syueth”, si è proceduto ad assegnare un peso prestabilito ad ogni emozione e sulla base delle emozioni rilevate si è calcolato il punteggio sentimentale di ogni commento.

Sono poi stati considerati come “negativi” i commenti aventi punteggio sentimentale pari o inferiore a 0 e come “positivi” tutti i commenti maggiori di 0.

Una volta separati commenti positivi e negativi, si è eseguito su ciascun gruppo di commenti la valutazione delle frequenze tramite metodo idf rispetto ai solo termini in forma di nomi e si sono valutati i primi cinque termini per ordine di frequenza decrescente nel caso Ford Focus.

L’output di questo processo è riportato nella **Tabella 4**.

Focus	1°	2°	3°	4°	5°
Positivo	Interior	Quality	Look	Handle	Transmission
Negativo	Noise	Transmission	Sync	Feature	Clutch

Tabella 4. Tabella delle cinque espressioni più utilizzate nei commenti positivi e negativi del modello Focus del marchio Ford, rilevate tramite metodo di ponderazione tf-idf ed ordinate per frequenza in ordine decrescente.

Quel che si è ricavato nella **Tabella 4** è l’elenco degli attributi del modello Ford Focus maggiormente citati nei commenti suddivisi in positivi e negativi.

La tabella rappresenta dunque un elenco ordinato delle caratteristiche che i consumatori associano maggiormente ai benefici ed i difetti del modello Ford Focus.

Da questi risultati si nota come il modello Ford Focus nel 2014 fosse caratterizzato da una buona qualità generale e qualità degli interni ma da gravi problemi associati alla trasmissione e al rumore.

Per convalidare queste ipotesi si è approfondita l’analisi facendo ricorso alla valutazione delle co-occorrenze, con l’obiettivo di comprenderne meglio la percezione del consumatore riguardo al modello.

Per impostare l’analisi delle co-occorrenze, si sono individuati tre aggettivi positivi e negativi più utilizzati all’interno dei commenti e si sono valutate le caratteristiche ad essi più associate.

Gli aggettivi così individuati sono stati: “good”, “great” e “nice” per gli aggettivi positivi e “bad”, “terrible” e “awful” per gli aggettivi negativi.

Si sono così valutate le co-occorrenze dei tre aggettivi sia positivi che negativi con altre espressioni contenute nel testo e sono stati estratti i primi due, tre o quattro termini con le frequenze congiunte più elevate.

Si è di fatto valutato quali caratteristiche dell’automobile sono state maggiormente associate ad aggettivi positivi e negativi all’interno dei commenti rilasciati dai consumatori.

Si sono inoltre valutate le co-occorrenze dei termini utilizzati nei testi con le espressioni “problem” ed “issue”, per valutare quali fossero le parti o le caratteristiche difettose di ogni modello secondo il consumatore.

I risultati della valutazione delle co-occorrenze sono stati sintetizzati all’interno della **Tabella 5.**

Modello	1°(e 2°) termine più usato	Car. Pos.	Car. Neg.	Problemi
Prius	Mpg // drive Great, average // fun	Acceleration Handle Drive	NA	NA
Focus	Transmission // noise Problem, issue, fix //wind, grinding	Quality Look Interior	Acceleration Transmission Noise Sync	Transmission Noise
Highlander	Seat Comfort, leather, support	Look Drive Interior	Noise Wheel	Vibration
Tucson	Seat // Drive //Price Comfort // Fun // Great	Control Price	Brake Quality	Audio Dealership

Tabella 5. Tabella contenente nella prima colonna i nomi dei modelli Prius, Focus, Highlander e Tucson. Nella seconda colonna i primi termini per frequenza idf ritrovati nei commenti e sotto questi le rispettive espressioni ad essi più associate. Nella terza e quarta colonna sono riportate rispettivamente le espressioni con co-occorrenza elevata rispetto ai termini “good”, “great” e “nice” ed i termini “bad”, “terrible” e “awful”. Nell’ultima colonna sono riportate le espressioni maggiormente associate ai termini “problem” ed “issue”.

I campi contrassegnati da Na indicano che non sono state ritrovate associazioni o co-occorrenze con frequenza superiore a dieci all'interno di tutti i commenti.

Il simbolo “//” ha il compito di suddividere i termini più utilizzati e le espressioni ad esso associate. Nel caso Prius, ad esempio, ad “Mpg” sono associate le espressioni “Great” e “Average”, mentre a “Drive” è associata l'espressione “Fun”.

La **Tabella 5** costituisce una rappresentazione schematica delle espressioni maggiormente utilizzate nelle recensioni e dei termini ad esse associati, le caratteristiche positive e negative principali ed i difetti principali riscontrati dai consumatori.

Rimanendo sul caso Focus si può notare come le prime espressioni per frequenza riportate dai consumatori siano Trasmissione e Rumore, a cui sono associati termini negativi e problematici.

Ad esempio, al termine “noise” sono associate le espressioni “wind” e “grinding” che ci indicano la tipologia di rumore che i consumatori lamentano di sentire.

Passando alle colonne successive si possono notare le caratteristiche positive e negative maggiormente associate al modello.

Secondo quanto riportato dai consumatori, il modello Ford Focus presenta una buona Estetica e una discreta Qualità generale.

Tuttavia, risultano essere presenti valutazioni negative riguardo alla Trasmissione e al Rumore a cui si aggiungono le espressioni Accelerazione e Sync.

Si può concludere che nel 2014 il modello Ford Focus ha sicuramente subito delle ripercussioni in termini di valutazioni, questo a causa dei problemi rilevati e percepiti dai consumatori rispetto al Rumore ed alla Trasmissione.

Dall'interpretazione della **Tabella 5** si possono individuare i pregi ed i difetti della Ford Focus espressi attraverso le parole dei consumatori.

Si è riusciti quindi a sintetizzare i contenuti dei commenti, andando a ricavare una valutazione degli attributi dell'automobile attraverso le parole dei consumatori.

Quest'analisi ha reso possibile individuare i principali difetti e vantaggi del modello Focus del marchio Ford, permettendoci di ottenere informazioni anche sui modelli “avversari”.

Le informazioni che si sono raccolte rappresentano una fonte inestimabile di Feedback rispetto alle caratteristiche dell'automobile, che possono permettere ad un'azienda di

ottenere una rappresentazione del posizionamento del proprio modello nella mente del consumatore.

4.2 È possibile partendo da commenti valutare quanto le percezioni del consumatore sulle caratteristiche di un'auto impattino sul voto totale?

Lo scopo di questo paragrafo è quello di comprendere quali siano le caratteristiche, riportate all'interno dei commenti, che influenzano maggiormente il punteggio finale assegnato al modello da parte dei consumatori.

Con punteggio finale si fa riferimento al rating rilasciato da ogni consumatore assieme alla recensione in forma di commento.

Se prima si è valutato quali sono le caratteristiche principali associate ad un modello, ora si valuta quanto la percezione di queste caratteristiche effettivamente impatti sulla valutazione finale.

Si vuole capire se è possibile, attraverso l'analisi testuale e sentimentale applicate ai commenti, comprendere l'importanza e la rilevanza che i consumatori associano ad alcuni degli attributi di un modello.

Per sviluppare quest'analisi si sono individuate le caratteristiche maggiormente discusse e commentate all'interno delle recensioni del modello Ford Focus negli 2012, 2013, 2014 e 2015, come ad esempio la qualità degli interni o il livello di comfort.

Successivamente, si sono estratti i sottotesti dei commenti contenenti queste caratteristiche e se ne sono calcolati i punteggi sentimentali.

Sulla base di questi punteggi sentimentali e della valutazione rilasciata dal consumatore si è proceduto a sviluppare un'analisi di regressione multipla.

Operativamente l'analisi parte dal dataset dei testi puliti rispetto al modello Ford Focus negli anni 2012, 2013, 2014, 2015.

Successivamente si sono individuate come caratteristiche da ricercare nell'analisi le espressioni dalla frequenza maggiore all'interno dei commenti della modello, ovvero: "Quality", "Look", "Interior", "Acceleration", "Transmission", "Noise" e "Sync".

Si è poi eseguita un'operazione di lemmatizzazione dei testi puliti che ha generato un nuovo dataset in cui ogni riga contiene una parola rilevata nei commenti ed il rispettivo commento di appartenenza.

In questa fase di scomposizione si elencano quindi nel dataset tutte le espressioni utilizzate

e le recensioni che le contenevano.

Il passaggio successivo è stato individuare tra i commenti quelli che contenevano congiuntamente tutte le caratteristiche alla base dell’analisi.

Tuttavia, i commenti contenenti tali caratteristiche in maniera congiunta sono stati 0, pertanto, si è ridotto il numero degli attributi ricercati.

La combinazione finale che è stata selezionata è formata dalle seguenti caratteristiche: “Look”, “Drive”, “Transmission” e “Comfort”; queste caratteristiche sono state ritrovate congiuntamente all’interno di circa ottanta commenti.

Una volta individuato il nuovo dataset di partenza, formato dagli ottanta commenti, si è proceduto ad effettuare l’analisi sentimentale.

Ogni commento è stato suddiviso in quattro sottotesti, in modo tale che ogni sottotesto comprendesse al suo interno una delle caratteristiche considerate.

Si è poi calcolato il punteggio sentimentale associato a ciascuno dei quattro sottotesti di ogni commento.

Infine, si è creato un dataset contenenti i punteggi sentimentali dei quattro sottotesti, il codice identificativo del commento di appartenenza ed il punteggio rilasciato dai consumatori in quel commento.

doc_id	Look_sent	Drive_sent	Transmission_sent	Comfort_sent	Rating
1	+2	+1	-3	0	3

Tabella 6. Tabella del primo rigo del dataset usato nell’analisi della regressione multipla. Nella prima colonna è riportato il codice identificativo relativo al commento considerato, le successive quattro colonne riportano il punteggio sentimentale dei sottotesti che contengono rispettivamente le espressioni: “look”, “drive”, “transmission” e “Comfort”. L’ultima colonna contiene il punteggio rilasciato dal consumatore.

Partendo dal dataset appena creato, di cui è riportata la prima riga in **Tabella 6**, è stato possibile sviluppare una regressione multipla, avente per variabili indipendenti le colonne Look_sent, Drive_sent, Transmission_sent e Comfort_Sent e per variabile dipendente la colonna Rating.

I risultati della regressione così formata sono riportati nella **Tabella 7**.

Caratteristica	Coefficiente	P-value
Look	0.874	0.097
Drive	0.620	0.236
Transmission	1.196	0.088
Comfort	0.739	0.125

Tabella 7. Output della regressione multipla avente per variabile dipendente il Rating e per variabili indipendenti le variabili Look_sent, Drive_sent, Transmission_sent e Comfort_Sent. Nella prima colonna sono riportate le quattro caratteristiche considerate. Nella seconda colonna si ritrovano i coefficienti associati ad ogni variabile indipendente. Nell'ultima colonna sono riportati i p-value associati ad ogni caratteristica.

Analizzando l'output si può subito notare come i p-value siano complessivamente elevati, generalmente superiori o pari a 0.1, questo probabilmente a causa delle dimensioni ridotte del dataset con cui si è andati ad eseguire l'analisi.

Infatti, i coefficienti associati a “Drive” e “Comfort” presentano p-value troppo elevati per essere considerati come significativi e quindi poter essere ritenuti affidabili all'interno di questa analisi.

Tuttavia, valutando i risultati ottenuti si può vedere come “Look” e “Transmission” siano le uniche variabili ad essere caratterizzate da p-value relativamente bassi, infatti il livello di affidabilità associato è superiore al 90%.

Si è deciso di accettare tale livello di affidabilità e quindi di considerare le due caratteristiche come significative per l'analisi.

I coefficienti associati a “Transmission” e “Look” sono rispettivamente 1.196 e 0.874, questo risultato dimostra quanto un incremento unitario del punteggio sentimentale associato alle singole espressioni possa far aumentare il punteggio medio.

Per “Transmission”, ad esempio, un incremento unitario del punteggio sentimentale associato a tale caratteristica comporterebbe un incremento pari a 1.196 della valutazione finale, questo se tutte le altre variabili dipendenti rimanessero costanti.

Dunque, attraverso quest'analisi si è potuto rilevare l'impatto che le caratteristiche “Transmission” e “Look” hanno sulla valutazione finale.

Queste informazioni affiancate a quanto visto nel paragrafo precedente mostrano come la valutazione generalmente negativa della “transmission” possa aver influenzato direttamente il punteggio finale rilasciato dai consumatori.

Si può concludere che, utilizzando le funzioni della sentiment analysis, è possibile organizzare un'analisi in grado di rilevare l'importanza ed il peso associato a determinate

caratteristiche del modello da parte dei consumatori.

Infatti, è stato possibile valutare quantitativamente l'impatto che un miglioramento o un peggioramento delle percezioni del consumatore rispetto alla "Transmission" ed al "Look" avrebbe sul punteggio finale nel caso Ford Focus.

Si è quindi riusciti ad individuare quali sono le caratteristiche dalla cui percezione dipende maggiormente l'assegnazione del punteggio finale e se ne è quantificato l'impatto effettivo su quest'ultimo.

Questa tipologia di analisi permetterebbe ad un'azienda di comprendere gli aspetti che stanno maggiormente a cuore ai consumatori e che maggiormente impattano sulle valutazioni del prodotto.

Dunque, sfruttando questo tipo di analisi, si permetterebbe al marketing di individuare gli aspetti cardine su cui incentrare campagne pubblicitarie o attività di perfezionamento del prodotto e di massimizzarne gli effetti di eventuali investimenti.

Inoltre, presa visione dell'output della regressione sulle percezioni delle caratteristiche del modello, il marketing manager potrebbe definire come nuovo obiettivo dell'attività di marketing un incremento unitario del punteggio sentimentale associato ad un certo attributo.

Permettendo così di valutare numericamente eventuali progressi o fallimenti nel raggiungimento di tale obiettivo.

Quel che si è sviluppato è uno strumento anche in grado di fornire feedback al team di sviluppo rispetto alla qualità dei materiali e delle tecnologie impiegate nella fabbricazione del modello.

Tutte queste informazioni vitali per l'azienda e per il successo del prodotto sono state estrapolate dai commenti, sfruttando conoscenze di analisi testuale e sentiment analysis.

Tuttavia, durante lo sviluppo di quest'analisi è stato riscontrato un problema importante. Questo problema riguarda la limitatezza quantitativa dei dati, più nello specifico il basso numero di commenti annuali riportati all'interno del dataset, che per costruzione conteneva non più di 200 commenti annuali.

Questa ridotta dimensione dei dati, assieme alla ricerca di commenti con quattro espressioni comuni al loro interno ha ridotto notevolmente sia le dimensioni del dataset che l'affidabilità stessa dell'analisi.

Sicuramente dataset di dimensioni maggiori garantirebbero una maggiore affidabilità e permetterebbero di sviluppare analisi più ampie anche in termini di caratteristiche considerate.

4.3 Ricorrendo all'utilizzo di funzioni di text mining e sentiment analysis è possibile comprendere le cause dell'andamento dei punteggi medi di un modello negli anni?

In questo paragrafo si vogliono ritrovare le caratteristiche che hanno dettato l'andamento della serie storica dei punteggi medi e sentimentali negli anni.

Nello specifico si vogliono ritrovare le caratteristiche positive e negative maggiormente riportate nei commenti di ogni anno, così da capire quali pregi dell'auto possono averne definito il successo o viceversa quali difetti possono averne causato il fallimento in termini di punteggi.

Nella sezione finale del paragrafo si riporteranno alcuni degli spot pubblicitari sviluppati per il modello Ford Focus negli anni, valutando le caratteristiche contenute nel messaggio.

Lo scopo di quest'ultima sezione è di osservare se l'azienda ha sviluppato una strategia di marketing sfruttando le informazioni contenute nelle recensioni.

4.3.1 Rilevamento punteggi medi

In questa analisi si sono presi in considerazione tutte le recensioni del modello Focus del marchio Ford, nell'intervallo di anni 2005-2018, andando a calcolare il valore medio annuale dei punteggi rilasciati dai consumatori.

Si è scelto di considerare il modello Focus del marchio Ford poiché tale modello risulta essere caratterizzato da un valore medio dei punteggi rilevati nell'intervallo considerato pari a 3,467 a cui è associata una varianza pari a 1,836 una delle più elevate tra tutti i modelli. Definito il modello oggetto di analisi si è rappresentata graficamente la serie storica dei punteggi medi annuali rilasciati dai consumatori.

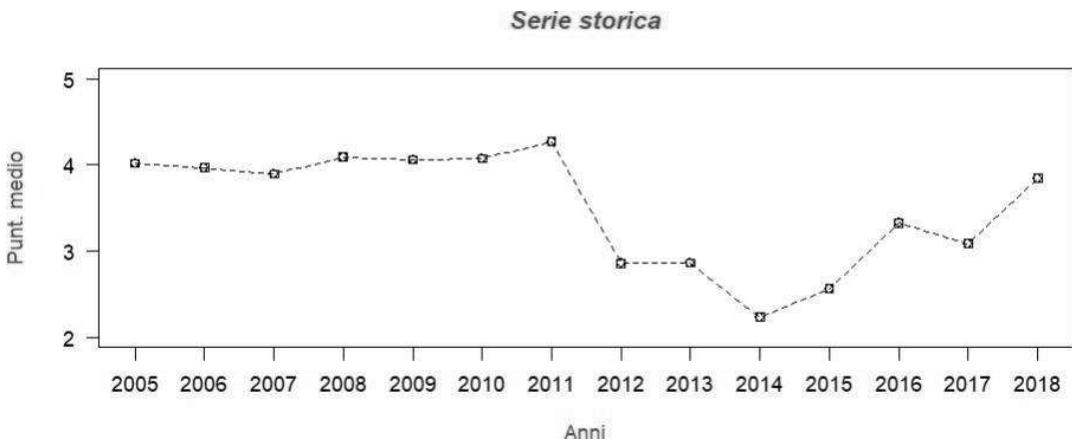


Figura 8. Grafico punteggio medio del modello Focus della Ford negli anni 2005-2018

L’obiettivo del grafico in **Figura 8** è quello di individuare le cause alla base dell’elevato valore della varianza, più nello specifico si sono voluti ritrovare gli anni in cui il punteggio medio presentasse picchi o crolli molto evidenti.

Dalla rappresentazione della serie storica, in **Figura 8**, è possibile notare un lungo periodo iniziale, composto dagli anni 2005-2011, durante il quale il marchio segue un andamento prevalentemente rettilineo, caratterizzato dal valore medio di poco superiore a 4.

Tuttavia, a questo periodo dalle elevate valutazioni per il marchio, segue un decisivo crollo del valore medio, che scende di più di un punto in un solo anno (2011-2012). Questo periodo di crollo delle valutazioni culmina nel 2014, anno in cui il modello registra il più basso valor medio delle valutazioni, pari a 2,3.

Da quest’anno in poi si può notare una ripresa del modello che termina nel 2018 in cui il punteggio medio del modello si avvicina ai punteggi rilevati negli anni 2005-2011.

Partendo da queste prime rivelazioni sorge spontaneo chiedersi cosa abbia causato questo crollo così drammatico dei punteggi del modello Ford Focus.

Tuttavia, essendo necessario lavorare sui commenti per ritrovare le cause della formazione di questa “conca” nei punteggi, si è ritenuto opportuno valutare se la curva dei punteggi sentimentali medi annuali associati ai commenti si comportasse in modo analogo a quella dei punteggi medi annuali.

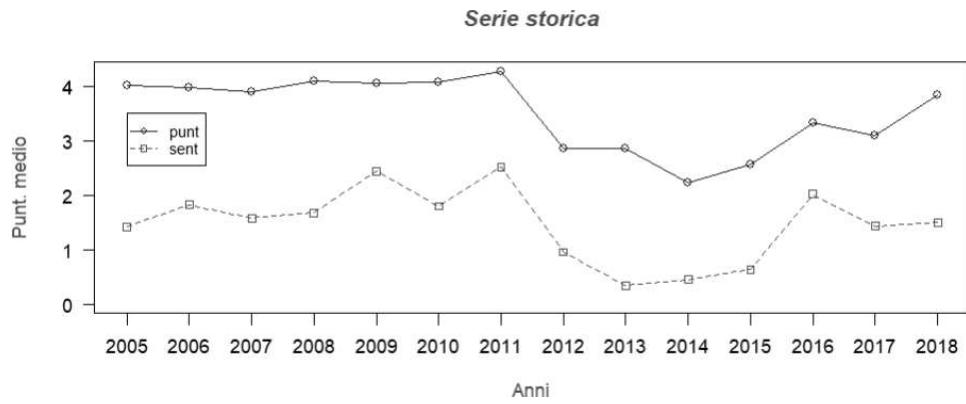


Figura 9. Grafico della serie storica dei punteggi del modello Ford Focus negli anni 2005-2018. La retta in blu rappresenta la curva dei punteggi medi rilasciati dai consumatori in ogni anno, mentre la retta rossa rappresenta la curva dei punteggi sentimentali medi annuali

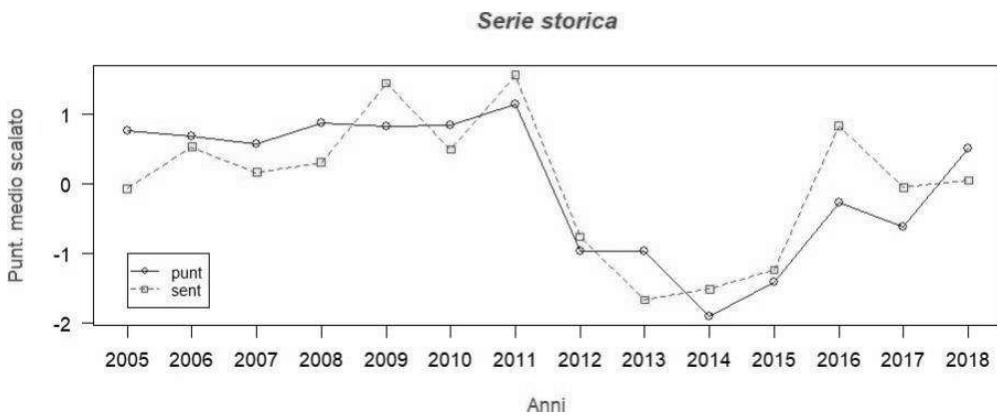


Figura 10. Grafico della serie storica dei punteggi scalati del modello Ford Focus negli anni 2005-2018. La retta in blu rappresenta la curva dei punteggi medi scalati rilasciati dai consumatori in ogni anno, mentre la retta rossa rappresenta la curva dei punteggi sentimentali medi scalati annuali.

Con i grafici in **Figura 9 e 10** si vuole valutare la presenza di un comportamento analogo tra la retta dei punteggi medi ed i punteggi sentimentali medi all'interno dell'intervallo di anni considerato.

Nel grafico in **Figura 9**, i punteggi sentimentali sono riportati nella loro scala originale, mentre, nella **Figura 10**, invece si è andati a scalare i punteggi sentimentali ed assegnati dai consumatori in modo da rendere visivamente più evidenti le differenze tra queste due serie storiche.

Da questi grafici, come ci si aspettava, si può notare come i punteggi sentimentali presentino un andamento analogo a quello dei punteggi assegnati dai consumatori.

Si nota quindi che il crollo del modello Ford Focus è rilevato sia dai commenti che dalle valutazioni.

4.3.2 Rilevamento ed analisi delle frequenze più utilizzate e degli spot pubblicitari

Valutata l'affidabilità dell'utilizzo dei commenti nello studio dell'andamento della curva dei punteggi medi si è passati ad impostare operativamente un'analisi delle espressioni positive e negative più utilizzate dai consumatori in ogni anno.

Questo processo ha il compito di studiare l'evoluzione e l'alternarsi delle caratteristiche considerate positive e negative da parte dei consumatori nei vari anni, con l'obiettivo di ritrovare le cause materiali che hanno portato all'abbassamento del valore medio dei punteggi.

Sfruttando il concetto di sottotesti, introdotto nell'analisi della domanda precedente, si sono suddivisi i commenti in sottotesti in modo da suddividere ogni commento in frasi dalla lunghezza limitata a poche parole.

Si è poi valutato il punteggio sentimentale associato ad ogni frase, sulla base del quale si sono suddivise le frasi in positive e negative.

Infine, si sono estratte con metodo di ponderazione idf le espressioni maggiormente utilizzate nell'insieme dei sottotesti positivi e di quelli negativi.

La scelta di calcolare i punteggi sentimentali associati ai sottotesti anziché agli interi commenti vuole ottimizzare ed affinare la ricerca delle caratteristiche positive e negative.

Nello specifico, si è voluto evitare il rischio di classificare erroneamente un'espressione sulla base del punteggio assegnato al commento.

Ad esempio, utilizzando i punteggi sentimentali associati ai commenti si rischia di classificare come positiva una caratteristica negativa come “noise” solo perché il punteggio complessivo del commento è positivo.

Suddividendo in blocchi il testo si è in grado di limitare questo problema e di rendere più affidabile il metodo di individuazione delle caratteristiche positive e negative.

Migliorato l'algoritmo per l'identificazione delle caratteristiche positive e negative maggiormente riportate nei commenti si è riportato nella **Tabella 11** l'output dell'analisi delle frequenze con ponderazione tf-idf.

Anno	Positivo	Negativo
2005	Drive, Fun, Interior	Transmission
2006	Comfort, Price, Look	Interior, Door
2007	Handle, Comfort	Tire
2008	Handle, Interior, Drive	Transmission, Tire
2009	Mpg, Price, Sync	Interior, Small
2010	Mpg, Sync	Interior, Steering
2011	Mpg, Drive, Handle	Interior, Transmission
2012	Interior, Speed	Transmission, Package, Back_seat
2013	Interior, Mpg, Handle	Transmission, Back_seat, Package
2014	Look, Comfort, Mpg	Transmission, Sync, Feature
2015	Interior, Feature, Quality, Sync	Speed, Small
2016	Interior, Package, Quality, Handle	Acceleration, Transmission, Shakes
2017	Comfort, Highway, Fun	Engine, Performance
2018	Comfort, Handle, Feel	Seat, Acceleration

Tabella 11 Tabella delle caratteristiche positive e negative più utilizzate in ogni anno, rilevate tramite metodo idf e con suddivisione tra commenti positivi e negativi basantesi sul calcolo dei punteggi sentimentali dei sottotesti. Nella prima colonna si ritrovano gli anni di riferimento. Nella seconda e terza colonna sono riportate rispettivamente le caratteristiche positive e negative dalla frequenza più elevata.

Nella **Tabella 11** è stato possibile riportare le caratteristiche che hanno caratterizzato il modello Ford Focus negli anni.

Considerando l'anno prima del drastico crollo dei punteggi, ovvero il 2011, si ritrovano “Mpg” “Drive” ed “Handle”, caratteristiche sempre presenti negli anni tra il 2005-2011. Si può notare come questi tre aspetti dell’automobile sono stati i più riportati tra gli aspetti positivi del modello Ford Focus negli anni caratterizzati dai punteggi più elevati, questo suggerisce l’importanza che i consumatori associano a questi aspetti.

È possibile affermare che “Mpg”, “Drive” ed “Handle” siano gli attributi che i clienti identificano come i punti di forza del modello e quindi aspetti che hanno sicuramente portato all’assegnazione dei punteggi medi elevati propri di quel periodo.

Nel 2011 queste caratteristiche “storiche” ritrovano ciascuna un posto tra gli attributi valutati come positivi, mentre “Transmission” ed “Interior” vengono valutati come negativi.

Nel 2012 le caratteristiche percepite come pregi e come difetti cambiano, in questa versione del modello Ford Focus ottiene un punteggio medio pari a 3 e tra le caratteristiche positive dell’automobile non si ritrovano più i termini storici “Mpg” “Drive” ed “Handle” ma bensì il termine “Interior” e “Speed”.

Nella versione del 2012 della Ford Focus si presentano anche nuovi problemi; infatti, alla trasmissione si aggiungeranno difetti o insoddisfazioni legate ai sedili posteriori ed ai “pacchetti”, varie forme di accessori di implementazione software come, ad esempio,

parcheggio automatico e guida assistita

Analizzando la **Tabella 11** si può affermare che la causa che ha determinato il crollo dei punteggi del modello Ford Focus nell'anno 2012 sia stata una rivoluzione del modello anche nei suoi attributi principali.

L'automobile è stata migliorata nella qualità degli interni e nelle prestazioni del motore, perdendo le percezioni di una macchina dalla buona efficienza in termini di chilometri per litro di benzina.

Si è inoltre introdotto un rinnovamento tecnologico tramite l'introduzione di pacchetti di servizi per l'assistenza alla guida, che non è stato ben visto dai consumatori che hanno riportato questi pacchetti tra le caratteristiche negative.

Terminata questa prima valutazione della tabella delle espressioni annuali più utilizzate, è sorta la curiosità di rilevare e studiare le operazioni di marketing che sono state adottate per garantire la ripresa del modello Ford Focus.

L'obiettivo di quest'analisi è quello di vedere come l'azienda Ford ha affrontato il crollo dei punteggi e di valutare l'attività di marketing svolta rispetto ai difetti che i consumatori hanno rilevato nell'automobile.

Nello specifico si vuole vedere se è possibile eseguire questa tipologia di dataset utilizzando i dati estratti dai commenti tramite text mining e sentiment analysis e valutando gli spot pubblicitari sviluppati negli anni considerati.

Da un punto di vista pratico sono stati presi in considerazione gli spot pubblicitari di tre anni differenti: 2008, 2012 e 2017; ricercandoli all'interno della piattaforma di contenuti multimediali “Youtube”.

I testi degli spot pubblicitari sono riportati nella **Tabella 12**

Anno	Messaggio pubblicitario
2008	<p>Un uomo disperso nella savana da anni viene ritrovato e reintrodotto nella società.</p> <p>Viene vestito, gli vengono acconciati i capelli e viene reso elegante, però nonostante l'apparenza l'uomo non smette di ruggire come un leone nella savana.</p> <p>Lo spot finisce con una voce fuori campo, facendo riferimento alla Ford focus, dice: "dicono che è cambiata del tutto, o quasi".</p>
2012	Design aerodinamico affascinante e che riduce i consumi, parcheggio automatico, prese d'aria dalla chiusura automatica per ridurre i consumi e mpg aumentato
2017	We all drive, some just for the fun of it

Tabella 12, Tabella dei contenuti riportati all'interno dei messaggi pubblicitari degli anni 2008, 2012, 2017. Nella prima colonna sono riportati gli anni di riferimento. Nella seconda colonna si ritrovano i contenuti degli spot.

Nell'anno 2008, come si vede nella **Tabella 12**, si presenta la prima importante rivoluzione del modello Ford Focus negli anni considerati.

Questa versione del modello presenta una maggiore attenzione al look ed al design degli interni, come si può vedere dalla **Tabella 11**.

Lo spot pubblicitario di quest'anno esalta le modifiche estetiche apportate al modello Ford Focus, andando a paragonare l'automobile ad un selvaggio che viene reso più elegante e che viene portato reso socialmente migliore.

La macchina viene resa esteticamente più apprezzabile e consona all'ambiente cittadino, senza perdere le "selvagge" prestazioni che la contraddistinguono.

Con questo messaggio l'azienda è riuscita efficientemente a rinnovare e modernizzare l'estetica del modello Ford Focus rassicurando i propri clienti rispetto a possibili preoccupazioni su cali di prestazioni e quindi su alterazioni riguardo alla meccanica del modello.

Nel 2012 viene introdotta una nuova rivoluzione che comporta una sfida maggiore, ovvero l'introduzione di funzioni e pacchetti software e di assistenza alla guida, che vanno ad automatizzare il modello.

Questo genere di innovazione viene introdotta con un approccio differente a quello utilizzato nello spot del 2008.

Anziché rassicurare il consumatore il messaggio che viene presentato ha lo scopo di sorprendere il cliente, elencando le caratteristiche innovative che il modello ha acquisito.

Questa strategia di marketing, assieme a possibili problemi tecnici legati alle nuove tecnologie, generalmente denominate sync, non hanno fatto aumentare i punteggi medi, anzi li hanno fatti crollare di più di un punto.

Le innovazioni tecnologiche introdotte, assieme a difetti associati alla nuova Trasmissione, hanno danneggiato i punteggi e l'immagine del prodotto anche negli anni successivi.

Consultando la **Tabella 11** si può vedere come negli anni 2015 e 2016 la tecnologia Sync, i vari pacchetti e le funzioni tecnologiche innovative siano passate negli gli attributi positivi dell'automobile, denotando un loro perfezionamento e quindi la risoluzione di eventuali difetti iniziali.

Infine, nel 2017 l'azienda riporta nello spot uno degli aspetti più importanti che era associato al modello negli anni 2005-2011, ovvero il “Fun to drive”, richiamando la piacevolezza dell'esperienza di guida.

Si può concludere che il marchio Ford ha sicuramente saputo sfruttare le informazioni contenute nei feedback dei propri clienti, sia per individuare i difetti rilevati dai clienti sia per ritrovare le caratteristiche da esaltare nell'attività di marketing.

4.3.3 Conclusioni

Si può concludere che grazie a quest'analisi è stato possibile individuare le caratteristiche positive e negative che hanno caratterizzato le valutazioni associate al modello negli anni 2005-2018.

Si è potuto valutare quali attributi hanno causato il crollo dei punteggi nell'anno 2012 e si è riusciti a risalire alle attività di marketing utilizzate dal marchio Ford per riprendersi dallo shock.

Sfruttando le funzioni di text mining e sentiment analysis è possibile per un'azienda ottenere varie informazioni sul proprio prodotto.

Nello specifico si rende possibile:

1. Individuare i benefici ed i difetti principali che i consumatori ritrovano nel prodotto.
2. Risalire alle caratteristiche dell'automobile che hanno portato ad un crollo

dei punteggi.

3. Rilevare la reazione dei consumatori difronte ad eventuali modifiche ed innovazioni del prodotto.

Inoltre, confrontando queste informazioni con gli spot pubblicitari utilizzati negli anni da altre azienda, un marchio sarebbe in grado di studiare le tattiche di marketing avversarie e imparare dagli errori delle altre aziende.

Nello specifico si è riuscito ad analizzare la strategia di marketing che è stata utilizzata per far fronte ad un'innovazione tecnologica, situazione che può essere difficile da affrontare per il marketing.

5. Conclusioni

L'obiettivo di questo elaborato era dimostrare ed esporre le potenzialità del text mining e della sentiment analysis, per quanto riguarda l'estrazione di informazioni utili ad un'analisi di mercato.

Tale obiettivo è stato concretamente rappresentato dalla formulazione di risposte alle domande portanti di ogni paragrafo del capitolo quattro, con lo scopo finale di ottenere una panoramica completa dell'universo del mondo delle automobili agli occhi dei consumatori, attraverso le loro stesse esperienze e valutazioni.

Partendo da commenti non guidati e valutazioni espresse su una scala da 0 a 5 si è riusciti ad impostare le seguenti analisi.

Si è stati in grado di risalire alle caratteristiche principali di modelli differenti, evidenziando i rispettivi punti di forza e debolezze, nonché i principali problemi, sviluppando un riassunto del mercato considerato.

Si è riuscito a valutare l'impatto che il punteggio sentimentale associato ad ogni caratteristica ha avuto sulla valutazione finale del modello, andando a identificare le caratteristiche più rilevanti e più influenti.

Infine, si è analizzata la storia di un modello dal punto di vista del consumatore e dei suoi bisogni, andando a valutare le caratteristiche che ne hanno definito il successo o il declino.

Tutte queste analisi sono state svolte facendo utilizzo del software Rstudio, tramite il quali si sono sviluppati codici in linguaggio R con funzioni di Sentiment analysis e text mining, andando a lavorare su recensioni e commenti.