Document Level Sentiment Analysis from News Articles [1]

Vishal S. Shirsat, Rajkumar S. Jagdale and S. N. Deshmukh

Case Study presented by:

Avanti Bhandarkar (BTech EXTC C008)
Chetan Popli (BTech EXTC C032)
Roshan Srivastava (MBATech EXTC J047)

Abstract

- Huge amount data generated on the internet; important to extract information. Data mining techniques used to solve diverse types of problems.

- In the era of News and blogs, there is need to extract news and need to analyze to determine opinion of that news reviews.

- Sentiment analysis finds an opinion i.e. positive or negative about particular subject. Negation is a very common morphological creation that affects polarity and therefore, needs to be taken into reflection in sentiment analysis.

- Proposed system uses online news databases from one resources namely BBC news. Three subtasks executed: categorizing the objective, separation of good and bad news content, data cleaning to get only what is required for analysis, steps like tokenization, stop word removal etc.

Introduction

Sentiment analysis

- Application of Computational Linguistics and Natural Language Processing which deals with the extraction of subjective information from text
- Used to determine the contextual polarity of a source of written or spoken media
- Data sources include texts, tweets, blogs, social media, news articles, product comments and reviews etc.

Levels of sentiment analysis

- Document level
- Sentence level
- Entity and aspect level

Proposed Approach

Dataset used

The BBC news dataset [2] was used in this paper, which included articles between 2014 - 2015 in the .txt format.

It contains 2225 documents which were split into 1490 articles for training and 735 articles for testing.

The categorical division of articles was as follows:

Category	No. of articles
Business	510
Entertainment	401
Politics	417
Sports	511
Tech	401

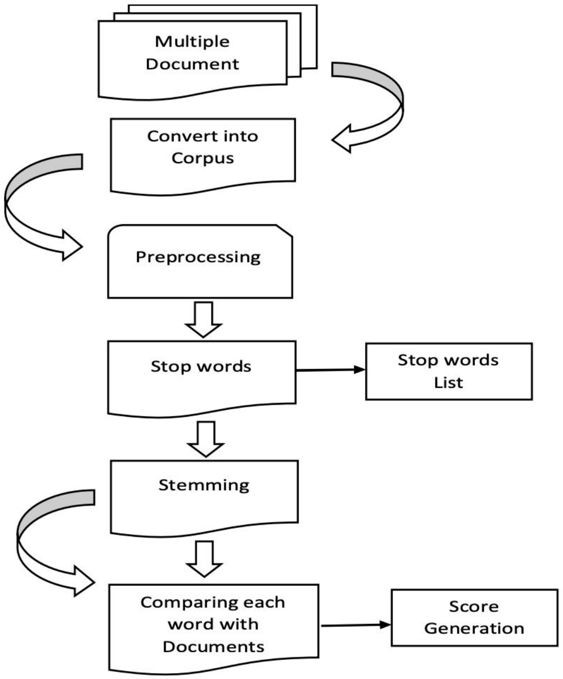


Figure 1: System diagram of the proposed approach

Pre-processing

Online data may contain irrelevant text such as HTML tags, scripts and advertisements

Pre-processing helps to clean the data, increase data sparsity and shrink the feature space

In this paper, pre-processing includes:

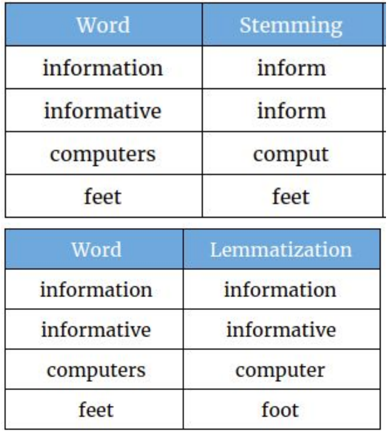
- removal of website URLs, hyperlinks, stop words, punctuation
- stripping of white spaces and numbers

Stemming and Lemmatization

Reducing a word to its word-stem or its root (known as lemma)

An example from this paper is:

‘The politician’s party is divided’ → ‘politician party be divide’



Word	Stemming
information	inform
informative	inform
computers	comput
feet	feet

Word	Lemmatization
information	information
informative	informative
computers	computer
feet	foot

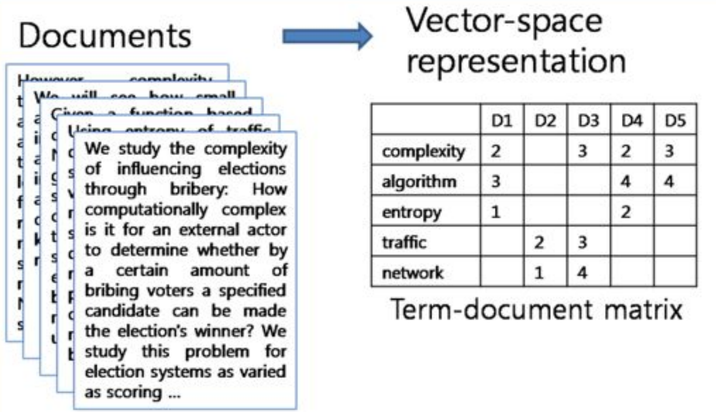
Figure 2: Stemming (top) and lemmatization (bottom)

Term Document Matrix (TDM) Generation

A Term Document Matrix describes the frequency of terms occurring in the processed dataset.

Rows = relevant terms
Columns = articles in the corpus

Thus, an entry (i, j) in a TDM represents the frequency of term i in article j



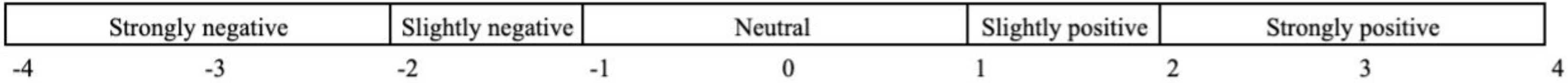
We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 3: Vector - space representation of documents in the form of a Term Document Matrices

Sentiment Analysis

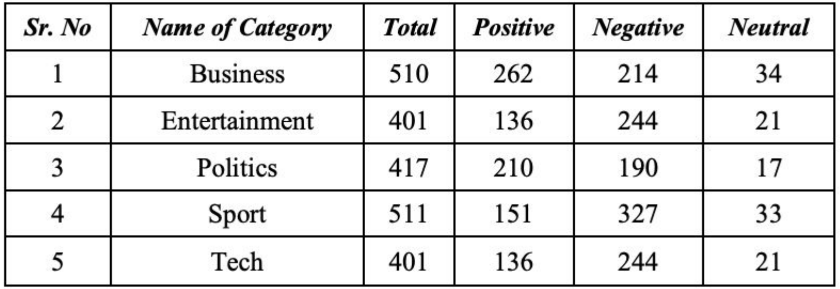
The paper used lexicon-based sentiment analysis; VADER [3] used to assign polarity between -4 and 4 to each word as follows:



The average of polarities in an article determined the overall sentiment associated with the article.

Results

Table 1: Sentiment analysis of documents



Sr. No	*Name of Category*	*Total*	*Positive*	*Negative*	*Neutral*
1	Business	510	262	214	34
2	Entertainment	401	136	244	21
3	Politics	417	210	190	17
4	Sport	511	151	327	33
5	Tech	401	136	244	21

Table 1 shows a categorical breakdown of articles along with the overall sentiments (positive, negative or neutral) associated with the article.

The proposed system applies a sentiment score function to the document based on the analysis of individual sentences.

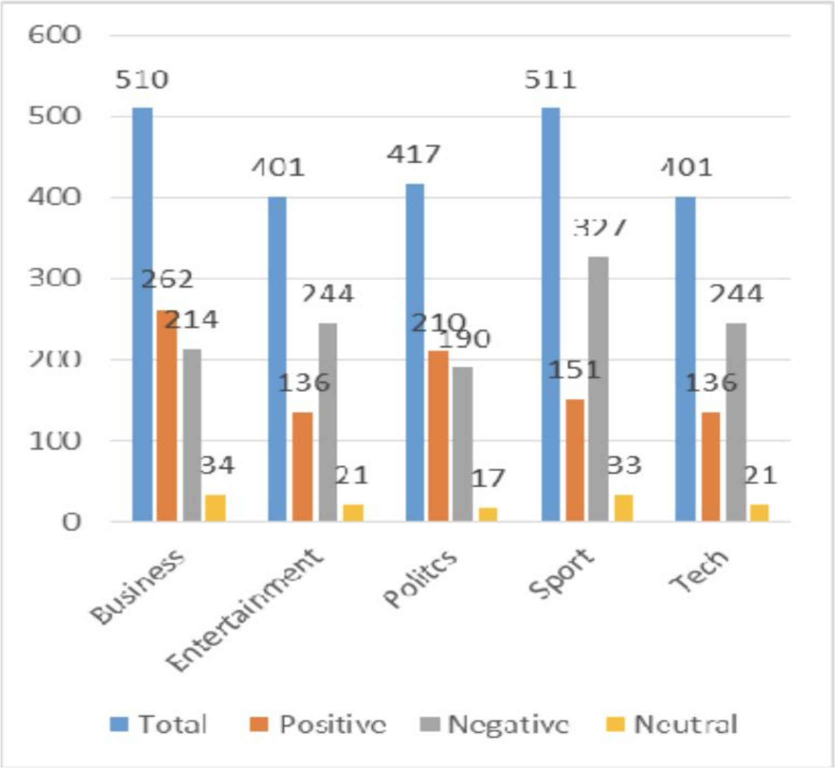


Figure 2: Graphical representation of the results

Conclusion and Disadvantages of the Proposed System

The Entertainment and Tech categories have an equal number of articles; their polarity distribution is also the same with a higher number of negative articles as compared to positive articles.

On the other hand, the Business and Sport categories also have an equal number of categories but polarities displayed across these categories are very different. Business articles show positive polarity while sports articles have a strong negative polarity.

However, these results may not be an accurate representation of document-level polarities for the following reasons:

- The proposed system is too simple to deal with complexities of language within news articles such as word ambiguities or negations

- Some of the articles show multipolarity i.e. the author displays a variety of opinions across a sentence which makes it difficult to accurately categorise sentences and in turn article

- The proposed system takes words at face value instead of adding context and gives poor results for articles containing sarcasm or irony

Potential Solutions to Disadvantages

Word ambiguities

Three common ambiguities are lexical, semantic and syntactic ambiguity.

Solutions:

- Part of Speech (POS) tagging
- Word embedding generation (Word2Vec, GloVe, Context2Vec etc.)
- Sense semantic relations between words by using a model like WordNet

Negations

Negations in sentences may affect sentence polarity based on kind of negation and context of use.

Solutions [4]:

- Heuristics based on Part of Speech (POS) where negation appears
- Analyse negations in conjugations and define exceptions where negations to not affect the sentence
- Take diminishers and morphological negators into consideration

Multipolarity

Multipolarity is when the text of interest contains varying sentiment, which may lead to a confusing overall sentiment of the text.

Solution:

- Determine words of interest at a sentence level and decide polarities associated with them; apply the Chi square test to understand level of multipolarity [5]

Sarcasm

Solutions:

- Use word embeddings to find sentiment incongruity [6] (usually a good determiner for sarcasm)
  For e.g. 'I love dying in this game!' is sarcastic despite the word love have strongly positive polarity

- Take numerical sarcasm into consideration [7] and compare it with adjectives or descriptors in the sentence
  For e.g 'He drove so slowly, only 100 km/hr' shows a clear contradiction between the numerical value and the adjective slow

Advantages and Applications of the Proposed System

The advantages of the proposed system are as follows:

- Application of pre-processing before the analysis stage transforms data into a more digestible form for determining polarity and associated sentiments
- Multivariate analysis of the generated Term Document Matrix reveals main themes and topics which help with accurate sentiment analysis
- Use of VADER allowed the authors to quantify how positive or negative a sentiment is; it also helped them incorporate slang and colloquial aspects of language

The potential applications of the proposed system are as follows:

- Understand topics and sentiments which are preferred by readers by correlating trends in in news articles in various categories with their associated sentiments
- Decide whether an article has any bias from the author's side, especially in case of political or controversial topics (can be achieved by modifying the polarity association)

References

[1] Shirsat, Vishal S., Rajkumar S. Jagdale, and S. N. Deshmukh. "Document level sentiment analysis from news articles." 2017 international conference on computing, Communication, Control and Automation (ICCUBEA). IEEE, 2017.

[2] Greene, Derek, and Pádraig Cunningham. "Practical solutions to the problem of diagonal dominance in kernel document clustering." *Proceedings of the 23rd international conference on Machine learning*. 2006.

[3] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. No. 1. 2014.

[4] Farooq, Umar, et al. "Negation Handling in Sentiment Analysis at Sentence Level." *J. Comput.* 12.5 (2017): 470-478.

[5] Marchand, Morgane, et al. "[LVIC-LIMSI]: Using Syntactic Features and Multi-polarity Words for Sentiment Analysis in Twitter." *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013.

[6] Camp, Elisabeth. "Sarcasm, pretense, and the semantics/pragmatics distinction." *Noûs* 46.4 (2012): 587-634.

[7] Kumar, Lakshya, Arpan Somani, and Pushpak Bhattacharyya. "" Having 2 hours to write a paper is fun!": Detecting Sarcasm in Numerical Portions of Text." *arXiv preprint arXiv:1709.01950* (2017).

Thank You!