**Comments Mining With TF-IDF: The Inherent Bias and Its Removal**

**Abstract:**

Text mining have gained great momentum in recent years, with user-generated content becoming widely available. One key use is comment mining, with much attention being given to sentiment analysis and opinion mining. An essential step in the process of comment mining is text pre-processing; a step in which each linguistic term is assigned with a weight that commonly increases with its appearance in the studied text, yet is offset by the frequency of the term in the domain of interest. A common practice is to use the well-known tf-idf formula to compute these weights. This paper reveals the bias introduced by between-participants’ discourse to the study of comments in social media, and proposes an adjustment. We find that content extracted from discourse is often highly correlated, resulting in dependency structures between observations in the study, thus introducing a statistical bias. Ignoring this bias can manifest in a non-robust analysis at best and can lead to an entirely wrong conclusion at worst. We propose an adjustment to tf-idf that accounts for this bias. We illustrate the effects of both the bias and correction with seven Facebook fan pages data, covering different domains, including news, finance, politics, sport, shopping, and entertainment.

**Architecture:**

****

**Introduction:**

SOCIAL media and in particular social networks (SNS) are today’s major form of communication used on a daily basis. SNS platforms serve individuals and organizations that utilize these platforms to spread information. In essence, a great portion of interpersonal communication has shifted, and is still shifting, to online social platforms. The availability of such rich online communication data provides fertile ground for researchers across multiple disciplines. The analysis of messages in social media falls squarely within the domain of data science, and effective processing and analysis of those messages will have a growing impact on business and business Information Technology (IT) management and Information Systems (IS). As Saar-Tsechanskyarticulates “IS data science is uniquely positioned to focus on challenges at the nexus of data science, business, and society. Insightful observations on the shortcomings of state-of-the-art methods to address old and new business challenges can give rise to new data science problems and research that addresses them”. Our work highlights one such shortcoming – the popular use of standard text analysis techniques (specifically, tf-idf, as we later describe), applied to comment classification, resulting in biased analysis – and proposes an adjustment. In qualitative research, discourse is defined as ’anything beyond the sentence’, which refers to the way in which people integrate language with other forms of self-expression (denoted ’non-language’), such as acting, interacting, and body language. In computer-mediated discourse (CMD1) such as on SNS, ’language’ is replaced by textual messages, and ’non-language’ is replaced by writing style, such as deleting subject pronouns, using abbreviations, and adding signs and symbol. An interesting aspect of comment analysis with respect to the relationship between commenters, is the between-participants’ discourse, and in particular, the language and non-language expressions they choose while participating in an ongoing discussion. We discuss the between-participants’ discourse in more detail later in the paper.

**Existing system:**

A discussion on the limitations of text analysis when applied to computer-mediated discourse (CMD) was presented, in which they claim that existing text analysis systems are concerned with topic modeling (assigning topics to documents). In CMD data, however, text features are often either overlooked or manually extracted. Arazy and Woo highlight the importance of effective statistical natural language (SNLP) methods to the IS community of researchers and practitioners. Their focus, on collocation indexing for compound terms, points to tf-idf as “the de facto standard” for weighting schemes that take into account a local, document-specific factor, and a global corpus-level factor. After showing how standard tf-idf is ineffective for collocations, they proceeded to develop an adaptation. We have identified a similar challenge to tf-idf showing how its unadjusted use in the analysis of social media comments is inappropriate and may result in significant bias. In the recent IS literature we observe a semi-manual approach, where classification relies on user (often expert) generated dictionaries or user-defined textual features. The need for developing customized automatedmethods is thus called for.The dataset for the case study contains 48 UMD articles, with an aggregate total of 3,538 annotated comments, out of which 149 (4.21%) were labeled as PB comments. The task of this study is to classify comments as PB comments or Non privacy breaching (NPB) comments, using comment classification techniques. This task falls under the umbrella of sentiment analysis, as PB comments are a users’ way to express that they know more than the article reveals. Note that the classification method should be able to classify comments on both existing articles and new (future) articles.

**Proposed system:**

We proposed an adjustment that corrects for the discourse bias when applying tf-idf to comments to online posts. However, it should be noted that the main contribution of this paper is in exposing the potential misuse of tf-idf when applied to dependent text. Other approaches can also be considered to quantify and/or mitigate the bias. One example is that suggested by that surveys different sources as potential textual baselines. Another example is kfold cross-validation, which is commonly used in sentiment analysis, where each of the k samples contain a list of comments to different documents (unlike the common practice, which is random sampling). While k-fold is expected to give a fair solution to the bias, it may be prohibitively expensive when applied to large datasets. This limitation is crucial in the context of user-generated-content and social media, where Big Data is being considered. In our case studies, some comment threads exhibit high discourse correlation, whereas others don’t. Other domains may contain higher correlations. However, the adjustment presented in this paper is applicable to any level of correlation (even to a no-correlation scenario), as the adjustment coefficient is proportional to the observed bias. While we present prominent advantages of adjusting the tf-idf weights, there are some limitations as well. In particular, when term population is small, tf-idf weighting may be biased, yet the adjustment we introduce may overcompensate for this bias, introducing another bias. For instance, there are cases in which some terms are used only by a small group of people and therefore their bias correction could likely be flawed.

**Modules:**

* **Text preprocessing**
* **Enhancing classification with comments’ structure**
* **Articulating the bias and adjusting tf-idf**
* **Comparison of Spearman’s Rank correlation**

**Text preprocessing:**

The tf-idf weighting, first introduced in, stands for term frequency (tf ) inverse document frequency (idf ). Tf-idf weighting is commonly used in text mining and information retrieval to evaluate the importance of a linguistic term (commonly a unigram or bigram) in a studied corpus. Term importance (weight) increases with the term’s frequency in the text, yet is offset by the frequency of the term in the domain of interest (e.g., frequent words like ”the” or ”for” will be scaled down).

**Enhancing classification with comments’ structure:**

 Research on threaded discussions makes use of comment structure. A threaded discussion is characterized by the inherent relationship between the threaded comments, and the textual dependency between them [26]. The relationship between comments, both structural and lexical, is highly recognized as useful in classification as seen in [31] who study the quality of comments in a threaded blog discussion. On top of the standard unigrams (words’ tf-idf score) and conversational features such as comment length, order, and number of replies, they studied comments with respect to their lexical similarity to the preceding comment and following comment. Interestingly, they show that, for the purpose of comments’ quality, lexical similarity features are dominated by conversation features, and that unigrams are the least informative feature.

**Articulating the bias and adjusting tf-idf:**

Our second contribution is a proposed statistical correction to the bias, delivered by modifying tf-idf. Our statistical correction is kept simple for practical reasons, and alternative corrections are discussed. In contrast to the literature on threaded discussion where the focus is to leverage dependency to derive predictors, we ”flatten” the threads and remove the dependency, to extract information from the remaining unigrams. We believe that our approach will make the unbiased analysis of comments more accessible to the IS community.

**Comparison of Spearman’s Rank correlation:**

 Compare the Spearman’s Rank, and Quantile Rank correlations of the different approaches across the same Facebook pages examined with RMSD. Table 4 summarizes the difference in Spearman’s Rank correlation between the proposed adjusted tf-idf approach and the other two alternatives. Overall, the Rank correlation is fairly low (lower than 0.3), with a slight advantage to the k-fold CV approach compared with our adjusted approach, and to the adjusted tf-idf approach compared with the non-adjusted tf-idf approach. As the quantile increases, and higher values of tf-idf weights are considered, the proposed adjusted tf-idf approach outperforms the alternatives in terms of capturing the relative importance of terms in the data, and the k-fold CV approach becomes the least competitive (see Figure 6). This result is consistent across all Facebook pages, and the different quantiles, and significant at the 1% significance level.

**Algorithm:**

An algorithm in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends.



**Future work:**

This paper is concerned with the bias that CMD introduces to the word weights assigned by tf-idf for the purpose of comments classification. Ignoring this dependency-bias may lead to one of two possible outcomes: if the dependency is intentionally ignored (assumed to introduce no bias to the model), the analysis may not be generalizable to out-of-sample observations. This will become evident when the method is applied to a holdout dataset. However, if the dependency is unintentionally overlooked, and comments on the same document are randomly split between training and holdout sets, the same statistical bias will be repeated in both sets. As a result, the researcher may reach incorrect conclusions regarding his analysis. In this paper, we exemplify this observation via examples. We further show how adjusting tf-idf to account for this dependency can significantly reduce this bias. Tf-idf is widely used for text pre-processing and feature engineering, as can be seen in these recent examples. An increasing emphasis on short texts and dynamic corpuses is evident in the literature.

**Conclusion:**

Many contemporary social media web sites, with Facebook being one, incorporate commenting systems that allow people to respond to posts on the web sites. Commenting systems encourage discourse exchanges between participants. Between-participant discourse has been proven to be content-dependent. This dependency, as we have shown in this paper, introduces bias to tf-idf, a widely used term weighting technique used for text preprocessing. We have further shown that this bias, if ignored, can manifest in a non-robust method at best, and can lead to an entirely wrong conclusion as worst. For decades, tf-idf has played a major role in information retrieval and text mining. Early on, tf-idf aimed to model the relationship between a term and a document in reference to a corpus. Over the years, the main practical usage of tf-idf (and variations thereof) was to model topics, and match queries to topics. In recent years, tf-idf has been harnessed to the classification of sentiments of short sentences (e.g., tweets and comments) - a significant deviation from its intended usage. The short-sentence environment, and even more so the tweet and comment environments, introduce two major challenges to tf-idf. Firstly, a large reference document collection is typically unavailable. Secondly, there exists between-participants’ discourse. As a result, classical tf-idf generates biased weights when applied to sentiment analysis. In this study we discuss this bias, we show that it mainly relates to discourse and topic, and that it is not part of the sentiment, in particular when the goal is prediction (out-of-sample observation with new topics and new discourse). Despite limitations of our approach, it is evident that bias correction of the sort we propose can significantly improve comment classification accuracy and processing time. As we show in three different examples, bias indeed exists, hence our approach is very relevant. In fact, bias of the type we have identified in comment and tweet environments exists in other domains as well, as we have identified in yet unpublished research. This suggests that, prior to using tfidf, researchers better check for bias in their target domain, and apply corrective adjustments where applicable to avoid undesirable analysis results.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL