iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Identification And Classification Of Noise In Converting A Document In Web Page

Bisma Sultan[1], Lalitsen Sharma[2]
[1,2]Department of computer science & it, university of Jammu, J&K

Abstract*: World Wide Web made the internet popular and attracted individuals /organizations to easily exchange information. Many web authoring tools were developed to create web documents fast and WYSIWYG way. On the other hand every application of the office includes an option to save the document as web page. The compliance of standard becomes an issue here and as a result a noise is said to enter in the web page. The noise not only disturbs the presentation of the web page but also takes extra storage. This paper presents a study on MS word documents which were converted into web pages and tested run in four popular browsers viz. Google chrome, IE7, Mozilla Firefox and opera. The results show that web pages created with the different options of conversion leads to different forms of noise. Results also show that the noise arises because of different browsers responding differently to DOM.
Keyword: web page noise; MS word; DOM; HTML; word document.

I. INTRODUCTION

Since the birth of World Wide Web, internet has gained a lot of popularity. World has reached an era were Organizations/individuals communicate and exchange information through internet. A number of web authoring tools have been developed for creating web pages fast and in a WYSIWYG manner. On the other hand, MS word uses three options to save a document as web page. Whenever we need to put a Microsoft word document on a web, the word document should be first converted to HTML so as to display it on a web. MS Word offers three ways to convert a Word document into web page which are "web page", "web page, filtered" and "single file web page". When we save a word document as web page using "web page" type option of MS Word, MS Word adds Microsoft specific tags to format the web page so that we continue to edit our pages using full functionality of word. The size of the web page created using this type is smaller than the web pages created using "single file web page". In case of web page, filtered option only regular tags are used to format the web page. This option creates the smallest size web pages compared to other two types. When we open a web page which was saved as filtered HTML, we cannot use all the formatting features of MS Word to edit our content. In addition, "Web page" and "web page, filtered" format store HTML file and additional resource files separately. "Single file web page" format creates a single MHT file called web archive file. This web page also contains office specific tags to format the web page. Moreover, MS Word also adds some unnecessary HTML tags to these web pages. All these unnecessary, gratuitous, futile tags are called noise. This Noise takes away the look and feel of a web page, it adversely affects web data mining [1].

In this work various kinds of noise are identified using different web browsers. We used four popular web browsers namely Google chrome, Internet Explorer7, Mozilla Firefox and opera. After identifying different kinds of noises, we classified these noises into three different categories based on the source of the word document.

II. METHODOLOGY

In this work, 40 word documents were created using three different sources and then converted into web pages. The noise present in these web page noise is identified and then classified noise into different categories.

A. The three sources used to create a word document are:

1) We converted pdf's into word documents by using online pdf-to-word converter.
2) We created word documents by copying data from internet or any other source and pasting it in the MS Word.
3) We created word documents by simply writing the content in MS Word.

After creating these word documents, we then converted them into web pages by using three web page creation options of MS Word. The three formats provided by MS Word are "web page, "web page, filtered" and "single file web page". While converting a document into web page, MS word changes a 2 or 3 column page into a single page. It sets the value of position property of all images to absolute. A web page created using "web page" and "single file web page "option include font definitions, style

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

definitions and list definitions in the head tag. On the other hand, Web pages created using "web page, filtered" include only font and style definitions in head tag. "Single file web page" and "web page" type uses both <*img*> tag and VML for creating images. Web pages created using "web page filtered" format do not use VML for loading vector graphics on web page.

III.CLASSIFICATION

A. DOM Based Noise

When a browser receives raw HTML code from server, a structured overview of the HTML document is created which is called Document object Model. Each HTML page corresponds to a DOM tree in which tags represent internal nodes and the detailed texts, images or hyperlinks represent the leaf nodes. [2]. Web page noise arise because different browsers handle DOM differently. MS word presents the web page similar to the way the page would be displayed in Microsoft Internet Explorer [4]. When a word document is converted into web page using "single file web page" format, a single archive file is created. This file stores the HTML file and additional resource files in a single file which is called web archive file(*.mht). It is seen that MHT files cannot be opened in Google chrome and Mozilla Firefox. Images present in web pages created using "single file web page" format are not opened in opera which results in unnecessary spot and therefore represent noise. In addition, not all the tags added by MS word are supported by opera which results in noisy elements on the web page.

Figure a

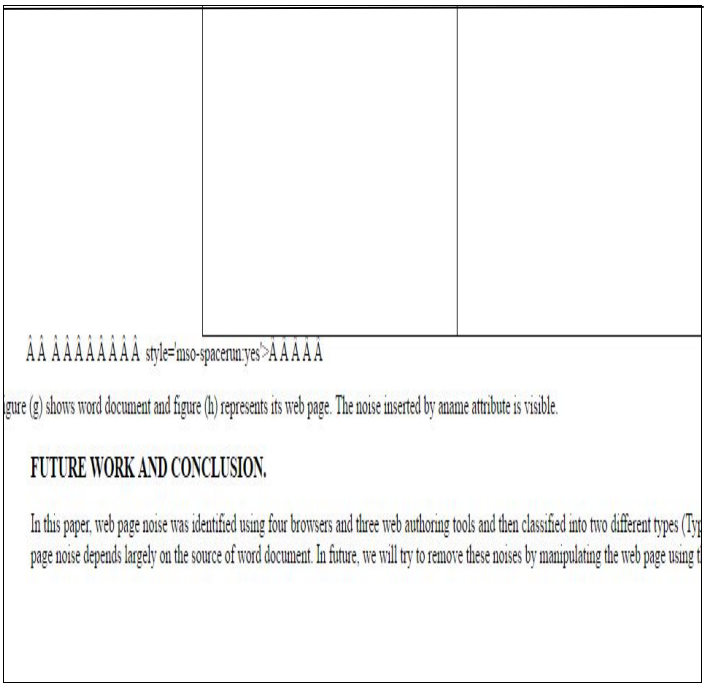


Figure b

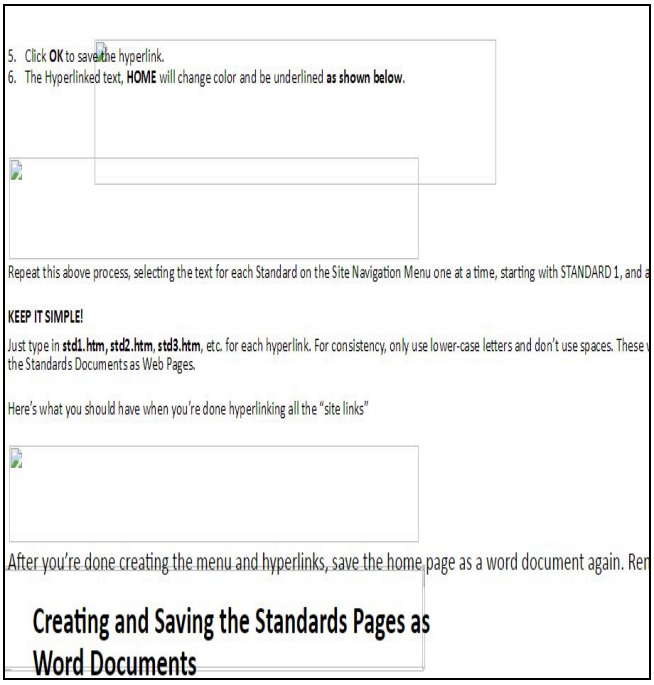


Fig. 1 Example of DOM Based Noise.

Fig 1 shows how the web pages created using "single file web page" option look when run on opera browser. . In figure a, Noise is inserted because of *"mso-spacerun"* style which is not supported by Opera and in figure b, noise is added by images present in the web page which did not open in opera browser. These images do not open in opera because MS word uses Vector Markup Language (VML) for creating vector graphics on web. VML is not supported by browsers like Google chrome, opera and firefox and therefore images created using VML do not open in these browsers which represent noise.

B. Type1 noise

When the web pages are created from the documents which are generated using pdf-to-word converter, the following noisy elements are present in the code associated with the web page:

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

1) Unnecessary, redundant <*div*> tags were seen in the code associated with these web pages. A separate div tag for a single line or a single word was present in the code. These unnecessary <*div*> tags result in noise and redundant data.
2) Unnecessary <*a name*> attributes were present in the code. The <*a name*> attribute specifies the name of the anchor tag. <*a name*> attribute is used to create a bookmark inside a document. These <a name> attributes were unnecessarily created in the web page.
3) Absolute value of the position property: The position property specifies the type of positioning method used for an element (static, relative, absolute or fixed) [3]. Any element that is defined after the element whose value is set as absolute will take its place and absolute element will go on a different space. Absolute element becomes floating element and it becomes relative to the outer (relative) parent. Thus position property when set as absolute disturbs the environment and makes it difficult to read the web page content. MS word sets the position property of all images as absolute.

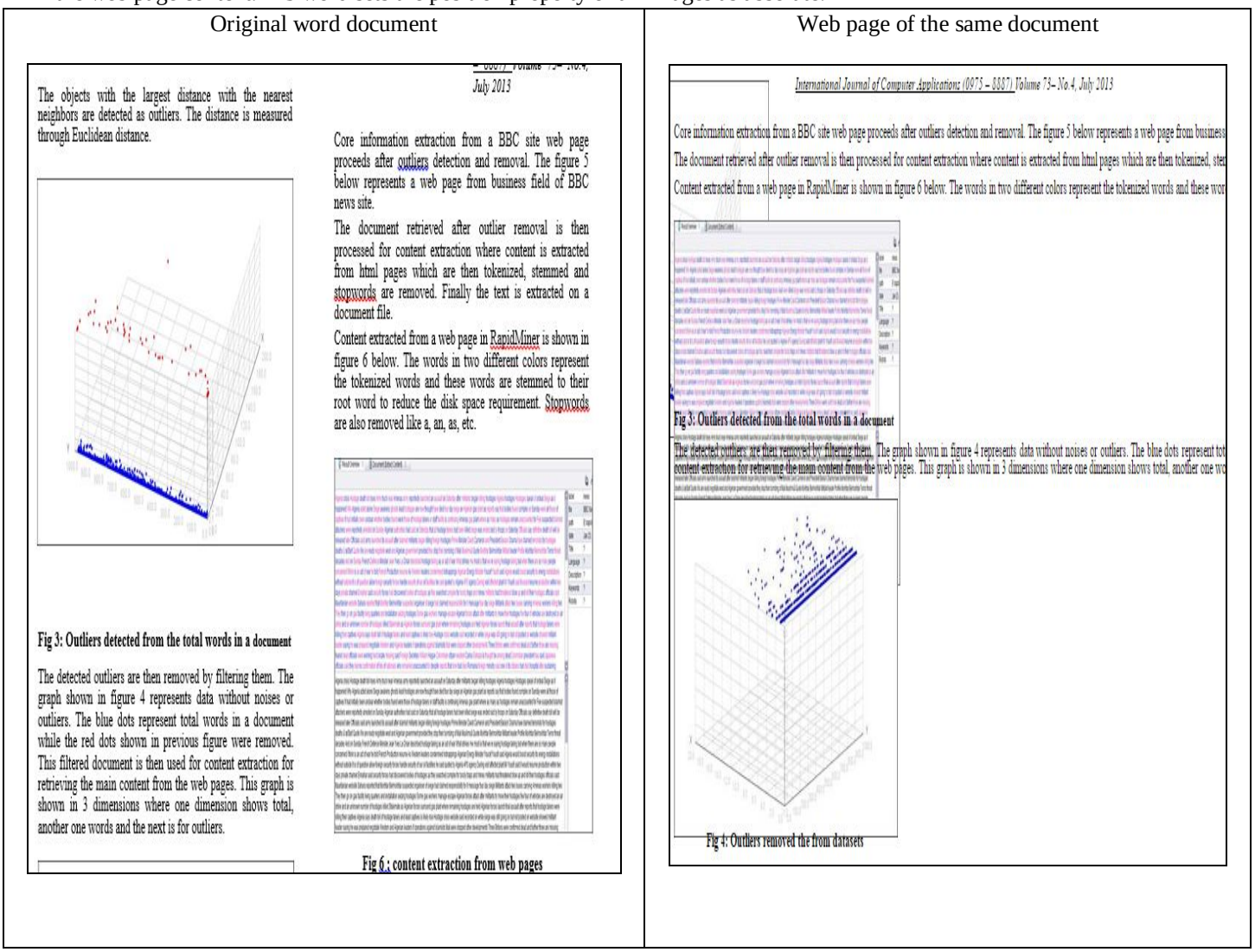
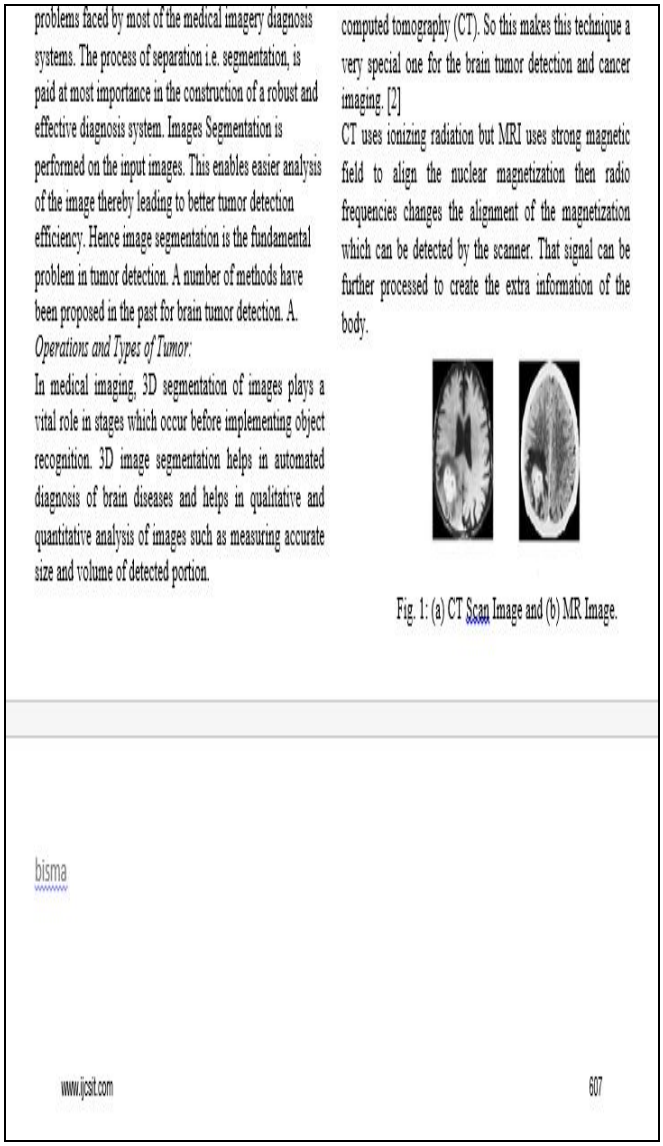


Fig.2 Example of Type 1 Noise.

Fig 2 represents a word document and the web page of the same document. It can be seen how noise affects the web page and makes it difficult to read its content. This noise belongs to Type1 where the value of position property is set as absolute which disturbs the environment and was seen when pages were run on all the four browsers.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Original word document


problems faced by most of the medical imagery diagnosis systems. The process of separation i.e. segmentation, is paid at most importance in the construction of a robust and effective diagnosis system. Images Segmentation is performed on the input images. This enables easier analysis of the image thereby leading to better tumor detection efficiency. Hence image segmentation is the fundamental problem in tumor detection. A number of methods have been proposed in the past for brain tumor detection. A.

Operations and Types of Tumor:

In medical imaging, 3D segmentation of images plays a vital role in stages which occur before implementing object recognition. 3D image segmentation helps in automated diagnosis of brain diseases and helps in qualitative and quantitative analysis of images such as measuring accurate size and volume of detected portion.

computed tomography (CT). So this makes this technique a very special one for the brain tumor detection and cancer imaging. [2]

CT uses ionizing radiation but MRI uses strong magnetic field to align the nuclear magnetization then radio frequencies changes the alignment of the magnetization which can be detected by the scanner. That signal can be further processed to create the extra information of the body.

Fig. 1: (a) CT Scan Image and (b) MR Image.

bisma

www.ijcst.com 607

Web page of the same document

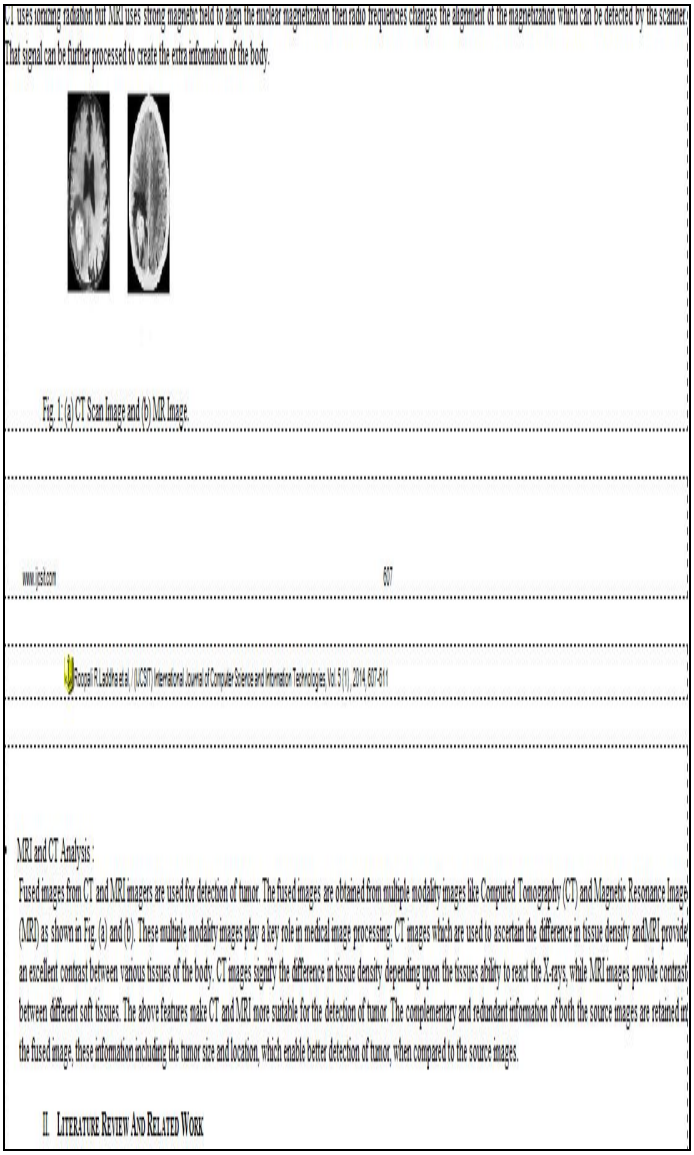

CT uses ionizing radiation but MRI uses strong magnetic field to align the nuclear magnetization then radio frequencies changes the alignment of the magnetization which can be detected by the scanner. That signal can be further processed to create the extra information of the body.

Fig. 1: (a) CT Scan Image and (b) MR Image.

www.ijcst.com 607

Bossi R. Ladde et al. / (IJCST) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 607-611

MRI and CT Analysis :

Fused images from CT and MRI images are used for detection of tumor. The fused images are obtained from multiple modality images like Computed Tomography (CT) and Magnetic Resonance Image (MRI) as shown in Fig. (a) and (b). These multiple modality images play a key role in medical image processing. CT images which are used to ascertain the difference in tissue density and MRI provide an excellent contrast between various tissues of the body. CT images signify the difference in tissue density depending upon the tissues ability to react the X-rays, while MRI images provide contrast between different soft tissues. The above features make CT and MRI more suitable for the detection of tumor. The complementary and redundant information of both the source images are retained in the fused image. these information including the tumor size and location, which enable better detection of tumor, when compared to the source images.

II. Literature Review And Related Work

Fig.3 Example of Type 1 Noise.

Fig 3 shows the original word document and the web page of the same document. From figure, it can be observed that web page contains unnecessary divisions which result in code redundancy. Here separate division has been created for a single word and a single line.

C. Type 2 noise

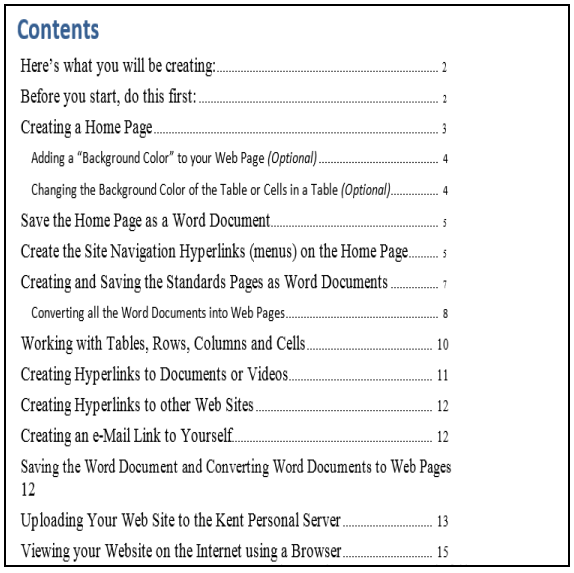
In case of web pages created from documents that are made by copy and paste, noisy elements are:

1) Unnecessary Hyperlinks*:* unnecessarily hyperlinks were present in web pages when run on all the four browsers.

2) Unnecessary <*a*> *name* attributes.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Original word document



Web page of the same document

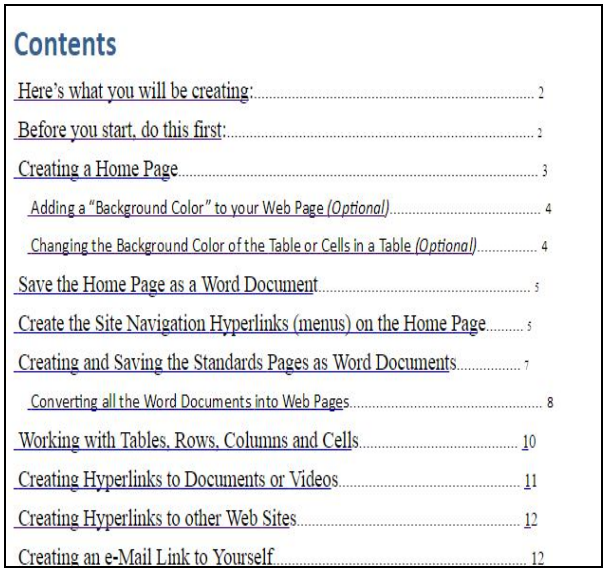
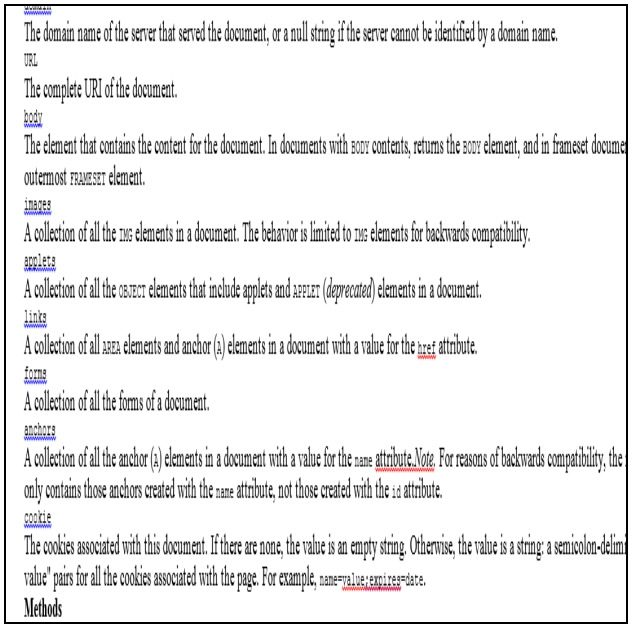


Fig.4 Example of Type 2 Noise.

Fig 4 shows a document containing text and the web page of the same document in which some text has been unnecessarily converted into hyperlinks.

Original word document



Web page of the same document

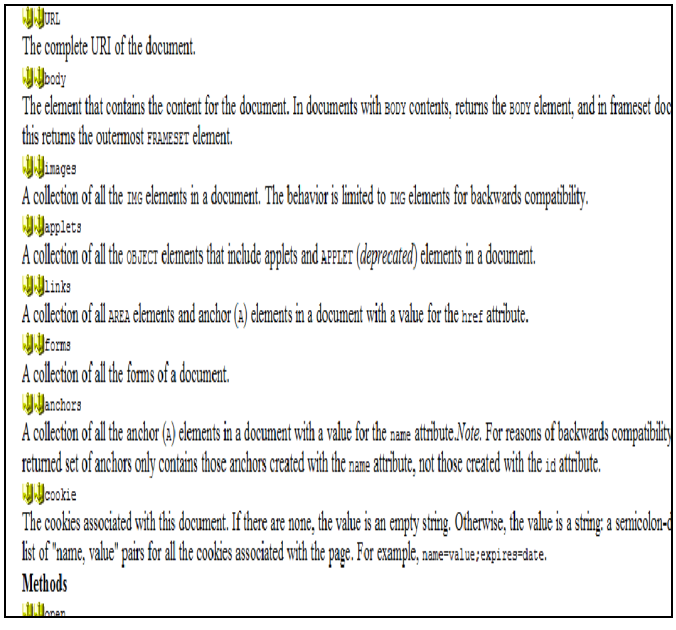


Fig.5 Example of Type 1 and Type 2 Noise.

Figure 5 represents a word document and its web page. The noise inserted by <*a name*> attribute is visible. This noise is present in both Type1 and Type2.

IV. RESULTS AND DISCUSSION

The experiment was conducted by converting around 40 word documents into web page using three web page creation options provided by MS word. These 40 word documents are then run on four browsers to identify noise. The analysis of code associated with these web pages show that MS word adds unnecessary, irrelevant tags which disturb the web page content and therefore are treated as noise. Moreover, web page noise also arise because various browsers treat DOM in a different way.

TABLE I
BROWSER SUPPORT

	Internet Explorer	Google chrome	opera	Mozilla firefox
Mso-spacerun	Yes	No	No	No
Tab-interval	Yes	No	No	No
Text-underline	Yes	No	No	No
Mso-tab-count	Yes	No	No	No

Table I lists four CSS properties used by MS word that are not supported by other browsers.

Mso-spacerun is used by MS word every time we use two spaces between sentences or words. That is whenever we use two space for separating words or Sentences, MS word creates something like "<spanstyle='mso-spacerun: yes'> ". >" This creates two spaces between the sentences or words. This style is not supported by other browsers and results in noisy elements when we run such pages on these browsers.

Tab-interval property allows authors to set the interval between default tab stops [5]. This property is not supported by other browsers and when we use tabs while creating a web page, MS word uses its specific attributes such as Mso-tab-count which are not recognized by other browsers and hence result in noise.

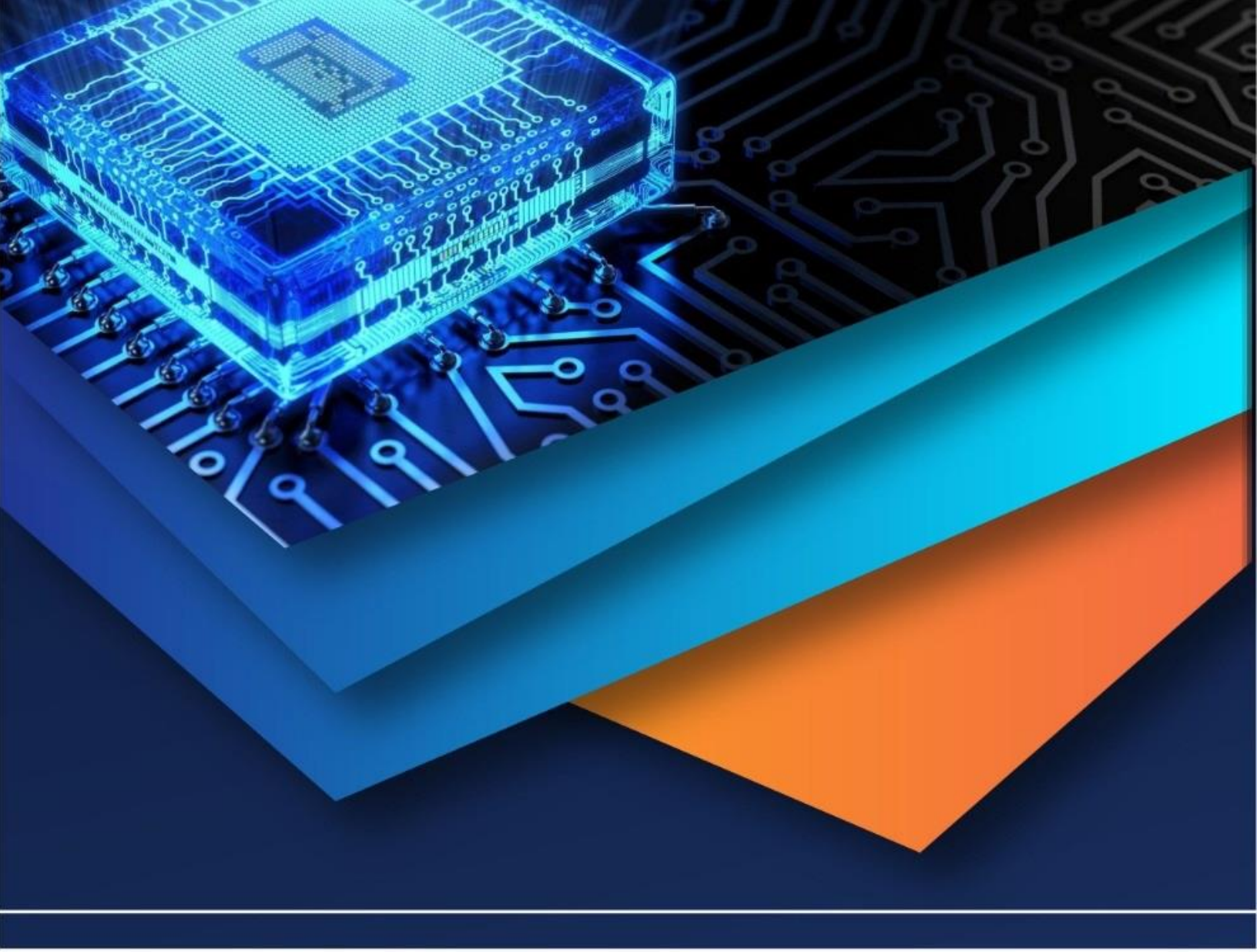
Text-underline property is used by MS word to specify position of underline decoration [6]. Its attributes are above (draws line above a text) | below (draws line below the text) | auto (default, decoration appears above the text). This property is not supported by Chrome, Firefox and opera.

V. CONCLUSION AND FUTURE WORK.

In this paper, noise that arises in web pages created using MS word is identified by analyzing the code associated with these web pages and that noise is then classified into different types. We used four browsers and three web authoring tools for noise identification. The experimental results show that noise arise because MS word adds some additional tags and properties which are not supported by other browsers resulting in noisy elements on web page. It also adds unnecessary, redundant tags which result code redundancy and make it difficult to read the contents of Web page hence are considered as noise. In future, we will try to remove this noise by manipulating the web page using the Document Object Model.

REFERENCES

[1] Anchal Garg and Bikrampal Kaur, "Web Page Performance Enhancement by Removing Noise," International Journal of Computer Applications (0975 – 8887) Volume 103 – No.6, October 2014.
[2] Lan Yi, Bing Liu and Xiaoli Li, "Eliminating Noisy Information in Web Pages for Data Mining," Proceedings of ninth ACM SIGKDD international conference on knowledge discovery and data mining.
[3] https://www.w3schools.com/cssref/pr_class_position.asp
[4] https://support.microsoft.com/en-us/help/212270/limitations-when-you-save-a-word-document-as-a-web-page
[5] https://www.w3.org/People/howcome/t/970224HTMLERB-CSS/WD-tabs-970117.html
[6] https://msdn.microsoft.com/en-us/library/ms531176(v=vs.85).aspx



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY