Literature Survey

1. Learning to Classify Documents According to Formal and Informal Style
https://journals.colorado.edu/index.php/lilt/article/view/1305/1137
Fadi Abu Sheikha and Diana Inkpen Published by CSLI Publications

It emphasizes the importance of vocabulary choice in determining style, with formal language characterized by longer words and Latin origin verbs, while informal language often includes phrasal verbs and idiomatic expressions.

The paper briefly introduces three machine learning algorithms used for classification: Decision Trees (DT), Naïve Bayes (NB), and Support Vector Machines (SVM).
It discusses related work in text classification by formality and genre, including efforts to classify texts by author, topic, and emotion.

The paper explains the main characteristics that differentiate formal and informal styles.
Characteristics of Informal Style include the use of the first and second person, contractions, abbreviations, phrasal verbs, subjective language, and slang.
Characteristics of Formal Style include the use of the third person, complex vocabulary, no contractions, appropriate and clear expressions, and objective language.

The authors built two datasets, one representing general-domain documents and another representing medical texts, to test the model's generalizability.
Features were extracted from the text, representing the main characteristics of both formal and informal styles.

The results are as follow:

Classification of General-Domain Texts (Document Level):
Three classifiers were used: Decision Trees, Support Vector Machine (SVM), and Naïve Bayes (NB).
Decision Trees achieved the highest accuracy, followed by SVM, with NB being the least accurate

Classification of General-Domain Texts (Sentence Level):
The same classifiers were used to classify sentences based on the same features used for document-level classification.
Decision Trees outperformed SVM and NB.
Similar to document-level classification, informal pronouns were the most significant feature, and phrasal verbs were the least significant.

Classification of Medical Texts (Document Level):
The study also classified medical texts using the same classifiers.
SVM achieved the highest performance in classifying medical texts, with Decision Trees also outperforming NB.

In all cases, the research employed 10-fold cross-validation to evaluate the classifiers, and F-measure and Accuracy were used as evaluation metrics. The findings suggest that the choice of classifier and features can significantly impact the performance of text classification, with Decision Trees and SVM generally outperforming Naïve Bayes, and specific linguistic features playing a crucial role in the classification process.

2. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa
https://www.sciencedirect.com/science/article/pii/S0306457321002375

J Briskilal , C.N. Subalalitha

In this paper, the authors attempt to classify idioms and literal expressions using ensemble techniques that combine two deep learning models, BERT and RoBERTa (a more robust version of BERT).

This paper attempts to classify idiomatic and literal sentences, using an ensemble method, with good accuracy. In sum, the main contributions of this paper are twofold: 1. Proposing an ensemble model to classify idioms and literal expressions using BERT and RoBERTa and 2. Constructing a dataset comprising 1470 idiomatic and literal sentences annotated by domain experts.

The authors compare the performance of their ensemble model with the baseline BERT and RoBERTa models. The ensemble model outperforms the baseline models in terms of F-score and accuracy, achieving a 2% improvement in accuracy.
The paper stresses on the importance of taking the weighted average of the BERT and RoBERTa models. The weighted average or weighted sum ensemble is a machine learning strategy that aggregates predictions from several models, with each model's contribution weighted according to its competence or performance.

Performance and robustness are the two primary reasons for using an ensemble over a baseline model. The proposed ensemble model is the reason for the reduced variance, increased accuracy and F-score.

3. No more beating about the bush: A Step towards Idiom Handling for Indian Language NLP
https://aclanthology.org/L18-1048.pdf
Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, Dipti Sharma

This paper presents IMIL, a multilingual parallel idiom dataset that contains 2208 commonly used idiomatic expressions in English, translated into seven Indian languages. Additionally, each idiom is annotated with sentiment information. The authors aim to demonstrate the usefulness of IMIL for two specific NLP tasks: Machine Translation and Sentiment Analysis.

For Machine Translation, the authors find that incorporating IMIL into translation models, particularly Phrase-based Statistical Machine Translation (SMT), results in significant improvements in translation quality, especially for sentences containing idiomatic expressions.

Regarding Sentiment Analysis, the authors show that IMIL contributes to improved sentiment analysis. This is achieved by integrating IMIL into the Stanford Sentiment Treebank. The sentiment annotations provided by IMIL help models capture the non-compositional nature of sentiments in idioms.

We can try to use this IMIL dataset in a separate experiment, for classifying sentences according to our criteria.

4. Idioms in Context: The IDIX Corpus
http://www.lrec-conf.org/proceedings/lrec2010/pdf/618_Paper.pdf
Caroline Sporleder, Linlin Li, Philip John Gorinski, Xaver Koch
Computational Linguistics / Cluster of Excellence MMCI
Saarland University

The paper introduces the "IDIX Corpus," which is a resource for studying idiomatic expressions in context. The paper focuses on annotating and distinguishing different usages of idioms, such as "literal," "non-literal," "mixed," "embedded in larger figurative context," "meta-linguistic," and "unclear/undecidable" usages, while also considering different senses of non-literal idioms.

The IDIX Corpus is designed to address the challenges of distinguishing between literal and idiomatic usages of expressions in context. By providing extensive annotations for idioms from the British National Corpus (BNC), this resource enables researchers to investigate and develop context-based methods for identifying idiomatic expressions in text.

It is particularly valuable for the following purposes:

Classifying Informal and Formal Texts: Informal texts often contain a higher density of idiomatic expressions, while formal texts tend to avoid them. By detecting idioms in a given text and analysing their usages, we can develop models and classifiers to distinguish between informal and formal texts.

Idiom Detection in NLP: Researchers can leverage the IDIX Corpus to improve idiomatic expression detection in NLP applications. This resource provides extensive annotations that can be used to train machine learning models to automatically identify idioms and differentiate between their literal and non-literal usages. This can enhance the accuracy of sentiment analysis, machine translation, and other NLP tasks.

5. A BERT-based Idiom Detection Model
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9869485

The paper titled "A BERT-based Idiom Detection Model" by Gihan Gamage, Daswin De Silva, Achini Adikari, and Damminda Alahakoon introduces a novel model for detecting idiomatic expressions within text. The proposed model aims to distinguish between idiomatic

and non-idiomatic expressions, ultimately improving the robustness and diversity of NLP techniques.

The key contributions of this paper can be summarized as follows:

Idiom Detection Model: The authors propose a model based on BERT (Bidirectional Encoder Representations from Transformers) that is fine-tuned using a token classification approach. This model is designed to classify words or tokens in a text as idiomatic or non-idiomatic.

Datasets: The authors evaluate their model using four different idiom datasets, including Idioment, EPIE (English Possible Idiomatic Expressions), and a dataset curated specifically for this study, theidioms.com. These datasets contain examples of idiomatic and non-idiomatic sentences in various forms.

Evaluation: The model is empirically evaluated on these datasets, and it achieves an accuracy of more than 94% across all experiments with 99.4 % accuracy on just the static idiom dataset from the EPIE corpus.

However, on inspecting the corpus, we note that as the idioms are static(always the same format) and repeated throughout the dataset, that the model built for this data particularly might be overfitting.
Therefore, we are currently trying to implement this paper ourselves.

Contextual Embeddings: The paper highlights the significance of contextual embeddings provided by BERT compared to traditional word embedding techniques like Bag of Words (BOW) and Word2Vec. Contextual embeddings capture the polysemy (multiple meanings) of words in context, making them suitable for detecting idiomatic expressions.

By classifying sentences or text segments as idiomatic or non-idiomatic, it can improve the accuracy of these applications and help users better comprehend the nuances and formality of the language used in different contexts.

6. Classifier Combination for Contextual Idiom Detection Without Labelled Data

Linlin Li and Caroline Sporleder

https://aclanthology.org/D09-1033.pdf

This paper discusses a two-stage approach for classifying idiomatic expressions .Idioms pose challenges for NLP systems due to their idiosyncrasies, such as violating selectional restrictions and subcategorization constraints.

We discuss their experimentation:

In the first stage, an unsupervised classifier determines idiomatic usage by analysing cohesive ties between the words in an expression and its context. It uses a measure of semantic relatedness to gauge these ties. The difference in connectivity between including and excluding the idiom's component words in a cohesion graph helps make this determination.

The second stage involves a supervised classifier that complements the unsupervised one. It is trained using a subset of the test data that the unsupervised classifier labels with high confidence. This two-stage approach allows for a more extensive feature set, improving classification accuracy, especially when cohesive ties are weak.

The text concludes by addressing the issue of imbalanced data and presents a solution involving iteratively enlarging the training set and boosting the literal class using non-canonical form variants of idioms (data augmentation) to maintain class balance.

The gist of the paper is emphasizing the importance of a cohesion-based approach using both unsupervised and supervised learning.

7. Automatic Idiom Identification in Wiktionary

Grace Muzny and Luke Zettlemoyer

https://aclanthology.org/D13-1145.pdf

In this research paper, the authors addressed the challenge of automatically identifying idiomatic phrases in language resources, with a primary focus on utilizing Wiktionary as a reference.

They collected a substantial dataset from Wiktionary(https://www.wiktionary.org/), comprising phrases, definitions, and example sentences. To automatically classify phrases as idiomatic or literal, they developed a machine learning model based on the Perceptron algorithm. Additionally, they engineered a set of lexical and graph-based features to aid in this classification process, taking into account synonym and antonym relationships, as well as graph distances between words in phrases and their definitions.

Their model significantly increased the number of identified idiomatic entries in Wiktionary, even when dealing with incomplete data. Moreover, they extended their approach to idiomatic phrase detection in example sentences, leveraging the Lesk word sense disambiguation algorithm(https://towardsdatascience.com/lesks-algorithm-a-method-for-word-sense-disambiguation-in-text-analytics-52c157a2fdff). This research offers a promising method for enhancing the understanding of idiomatic language and expanding idiomatic labels in dictionaries.

8. Translation of Idioms: A Hard Task for the Translator

https://www.researchgate.net/profile/Hossein-Vahid-Dastjerdi/publication/228837452_Translation_of_Idioms_A_Hard_Task_for_the_Translator/links/09e4150d54e9b79be6000000/Translation-of-Idioms-A-Hard-Task-for-the-Translator.pdf

The paper discusses the challenges of translating idiomatic expressions from one language to another. Idioms are linguistic expressions that are unique to a particular language and culture, making their direct translation difficult. The paper provides definitions and classifications of idioms, including colloquialisms, proverbs, slang, allusions, and phrasal verbs.

The primary focus of the paper is on strategies for translating idioms. The author suggests several approaches:

Using an Idiom of Similar Meaning and Form: This strategy involves finding an idiom in the target language that has both a similar meaning and similar lexical items to the source idiom.

Using an Idiom of Similar Meaning but Dissimilar Form: In this case, the meaning remains the same, but different lexical items are used in the translation.

Translation by Paraphrase: When no exact equivalent exists, the translator rephrases the idiom to convey its meaning, although the cultural and linguistic impact may be lost.

Translation by Omission: In situations where no equivalent or suitable paraphrase can be found, the idiom may be omitted from the translation.

Giving a Literal Translation: This strategy involves providing a word-for-word translation of the source language idiom, but it can lead to confusion if the target language readers are not familiar with the idiom.

The paper emphasizes the importance of maintaining the cultural significance of idiomatic expressions in translation. It suggests that, when possible, translators should aim to convey the intended effects and cultural nuances of idioms from the source language.

Note: this paper deals solely with the linguistic side to the problem and specifically on Arabic.

9. Classification of Idioms and Literals Using Support Vector Machine and Naïve Bayes Classifier

https://link.springer.com/chapter/10.1007/978-981-16-5078-9_42

This paper focuses on the classification of idiomatic phrases and literal phrases. The authors employ two machine learning algorithms, Support Vector Machine (SVM) and Naïve Bayes, to classify these types of phrases. They have created an in-house dataset comprising 1471 sentences, including 735 idiomatic and 735 literal sentences, which has been manually annotated by domain experts.

The experimentation involves data preprocessing techniques, such as removing duplicates, tokenization, stop word removal, case conversion, stemming, and lemmatization. Features are extracted, and the text is transformed into vectors using the TF-IDF method. The dataset is then divided into a training set (80%) and a test set (20%) for the classifiers.

The results of the experiments indicate that SVM outperforms Naïve Bayes, achieving an accuracy of 87.30%, while Naïve Bayes achieves 82.09% accuracy. The authors attribute the lower performance of Naïve Bayes to the small dataset and the probabilistic nature of the classifier.

10. Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification

https://aclanthology.org/2022.starsem-1.21.pdf

The paper discusses the task of idiom token classification, which involves distinguishing between the idiomatic and literal interpretations of linguistic expressions. It specifically explores how contextualized word embeddings from masked language models (MLMs), such as BERT, can aid in this classification. The authors propose a method to improve a BERT-based idiom token classifier by leveraging three types of embeddings: contextualized embeddings, uncontextualized token embeddings, and masked token embeddings. They conduct experiments on both English and Japanese datasets and show that these additional embeddings significantly enhance idiom token classification in a zero-shot setting.

The authors present a model called "BERT[vV; vN]" that leverages contextualized embeddings of the constituent words of a target phrase, achieving high accuracy in idiom token classification in a zero-shot setting.

The study demonstrates the effectiveness of using uncontextualized token embeddings and masked token embeddings in addition to contextualized embeddings. Combining these embeddings improves the classification accuracy further.

The authors conduct experiments on both English and Japanese datasets and show that their approach outperforms other baseline models, including SVM-based models and prior works that used contextualized embeddings.

They provide evidence supporting their assumption that when a target phrase is used in its literal sense, the uncontextualized and masked token embeddings tend to be similar to the standard BERT embeddings. This indicates the new embeddings' ability to capture idiomaticity.

11. Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions

https://aclanthology.org/2020.cogalex-1.9.pdf

In this research, the focus is on detecting idiomatic expressions and distinguishing them from literal interpretations, particularly in cases where the meaning depends on the context. The study leverages the ***Idiom Principle***, which suggests that idiomatic expressions are stored and retrieved as single units in memory. It also uses contextualized word embeddings (CWEs) such as Context2Vec and BERT to capture context-specific meanings of idiomatic expressions.

The experiments are divided into two settings: one using original pre-trained embeddings and the other using "Idiom-Principle-Inspired" models, where each idiomatic expression is treated as a single token. Results show that CWEs outperform non-contextualized embeddings, especially when the Idiom Principle is applied. CWEs can place potentially idiomatic expressions into distinct sense regions in the embedding space. These models can also provide suitable substitutes for ambiguous expressions in context, which is promising for tasks like text simplification.

The research uses the VNC-Tokens dataset to evaluate the models, achieving significant improvements in detecting idiomaticity in multiword expressions, especially in distinguishing literal and idiomatic senses. The Context2Vec model, inspired by the Idiom Principle, shows the most significant improvement, with robustness across different expressions. T-SNE (*t-distributed stochastic neighbour embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map*) plots reveal that CWEs, especially Context2Vec, are better at clustering idiomatic and literal senses in their semantic spaces.

The study further demonstrates the Context2Vec model's ability to suggest appropriate substitutes for removed idiomatic expressions in context, indicating a deeper understanding of the idiomatic sense in the model.